Dissertation
submitted to the
Combined Faculties for the Natural Sciences and for Mathematics
of the Ruperto-Carola University of Heidelberg, Germany
for the degree of
Doctor of Natural Sciences

presented by

M.Sc. Stephan Marius Tirier
born in: Essen, Germany
oral examination date: November 05, 2018

Dissecting tumor cell heterogeneity in 3D cell culture systems by combining imaging and next generation sequencing technologies

Referees: Prof. Dr. Roland Eils

Prof. Dr. Peter Angel

Für meine Eltern, Ursula und Christian Tirier,
meine Frau Stefanie Tirier, geb. Birnbach und unsere Tochter Marie

Acknowledgements

First, I would like to thank Dr. Christian Conrad and Prof. Dr. Roland Eils for their support and guidance during the last years and for giving me the opportunity to work in the fascinating and flowering field of single cell genomics. I am very grateful that I could use and further develop cutting edge technology to solve questions in biomedical research and to work in such a fruitful environment. In particular, I would like to thank Dr. Christian Conrad for great supervision, for giving me the freedom to explore and define directions of my PhD projects, for open and intense discussions and for his enthusiasm, creativity and energy that pushed me forward during the last years. Likewise, I would like to thank Prof. Dr. Roland Eils for mentoring, support and funding of my PhD work. Although I am not able to join the new research department in Berlin in the near future, I hope to keep close contact in the following years. In addition, I would like to thank Prof. Dr. Hanno Glimm and Dr. Christoph Merten for support and helpful comments during my thesis advisory committee meetings as well as Prof. Dr. Peter Angel, Dr. Sevin Turcan and Dr. Shirin Doroudgar for being members of the thesis defense committee.

I would especially like to thank Friedrich Preußer, a former master student, as well as my eilslabs co-workers Jeongbin Park, Dr. Zuguang Gu, Simon Steiger, Dr. Marcel Waschow, Björn Eismann, Foo Wei Ten, Dr. Teresa Krieger and Katharina Jechow for their contribution to the pheno-seq and the colorectal cancer project. I was very lucky to work with great collaboration partners in both projects and I would like to express my gratitude especially to Dr. Jan-Philipp Mallm, who was a great help during the first month and throughout my whole PhD work, as well as to Dr. Matthias Schlesner, Prof. Dr. Karsten Rippe, Prof. Dr. Christiane Fuchs, Prof. Dr. Fabian Theis, Lisa Amrhein, Prof. Dr. Hanno Glimm, Dr. Claudia Ball and Martina Zowada. Furthermore, I would also like to thank Teresa Krieger, Timo Trefzer, Lorenz Chua and Foo Wei Ten for proof-reading, the intelligent imaging group, the Synthetic Biology group and the whole eilslabs department for the great and friendly working atmosphere as well as for productive and fun group meetings and retreats. Many thanks also go to Dr. Dominik Niopek for helpful discussions and feedback. Importantly, I would like to thank the members of our department administration Manuela Schäfer, Corinna Sprengart, Dr. Jan Eufinger and Dr. Julia Ritzerfeld, who were helpful whenever necessary.

Finally, I would like to thank my family and my friends, especially my parents Ursula and Christian Tirier as well as my wife Stefanie for the support and trust during the last years.

Contributions

If not stated otherwise, all data presented in this thesis were obtained and analyzed by myself under supervision of **Dr. Christian Conrad** and **Prof. Dr. Roland Eils**. However, a significant amount of work, including method development/application and data analysis has been performed in collaboration with colleagues and I highly appreciate their great work that contributed to the success of the studies presented in this thesis. Thus, I will use the term 'we' throughout the whole work because I consider the acquisition of presented results as team-work. In the following section, I will specify all datasets that were obtained and/or analyzed in collaborative efforts. In addition, these contributions are also indicated in corresponding figure legends.

The following colleagues contributed to the pheno-seq project (section 2.1): Fluidigm C1 single cell libraries of MCF10CA cells were generated in collaboration with **Dr. Jan-Phlipp Mallm**. The HT-pheno-seq imaging protocol, the image processing pipeline and PhenoSelect were developed together with **Friedrich Preußer**. **Jeongbin Park, Simon Steiger** and **Dr. Zuguang Gu** performed RNA-sequencing data pre-processing and contributed to data analysis. **Friedrich Preußer** and **Dr. Marcel Waschow** analyzed imaging data and **Björn Eismann** performed light-sheet microscopy**. Prof. Dr. Hanno Glimm** and **Dr. Claudia Ball** provided the patient-derived colorectal cancer spheroid culture model and **Prof. Dr. Fabian Theis**, **Prof. Dr. Christiane Fuchs** and **Lisa Amrhein** adapted and applied the deconvolution approach based on maximum likelihood inference.

The following colleagues contributed to the colorectal cancer project (section 2.2): **Prof. Dr. Hanno Glimm, Dr. Claudia Ball** and **Martina Zowada** provided the patient-derived colorectal cancer spheroid culture models including associated metadata and whole exome DNA sequencing results. **Jeongbin Park** performed RNA-sequencing data pre-processing and downstream analysis was performed together with **Dr. Teresa Krieger**. The RNA-FISH image analysis pipeline was developed together with **Foo Wei Ten**.

Abstract

Three-dimensional (3D) *in vitro* cell culture systems have advanced the modeling of cellular processes in health and disease by reflecting physiological characteristics and architectural features of *in vivo* tissues. As a result, representative patient-derived 3D culture systems are emerging as advanced pre-clinical tumor models to support individualized therapy decisions. Beside the additional progress that has been achieved in molecular and pathological analyses towards personalized treatments, a remaining problem in both primary lesions and *in vitro* cultures is our limited understanding of functional tumor cell heterogeneity. This phenomenon is increasingly recognized as key driver of tumor progression and treatment resistance. Recent technological advances in next generation sequencing (NGS) have enabled unbiased identification of gene expression in low-input samples and single cells (scRNA-seq), thereby providing the basis to reveal cellular subtypes and drivers of cell state transitions. However, these methods generally require dissociation of tissues into single cell suspensions, which consequently leads to the loss of multicellular context. Thus, a direct or indirect combination of gene expression profiling with *in situ* microscopy is necessary for single cell analyses to precisely understand the association between complex cellular phenotypes and their underlying genetic programs.

In this thesis, I will present two complementing strategies based on combinations of NGS and microscopy to dissect tumor cell heterogeneity in 3D culture systems. First, I will describe the development and application of the new method 'pheno-seq' for integrated high-throughput imaging and transcriptomic profiling of clonal tumor spheroids derived from models of breast and colorectal cancer (CRC). By this approach, we revealed characteristic gene expression that is associated with heterogeneous invasive and proliferative behavior, identified transcriptional regulators that are missed by scRNA-seq, linked visual phenotypes and associated transcriptional signatures to inhibitor response and inferred single-cell regulatory states by deconvolution. Second, by applying scRNA-seq to 12 patient-derived CRC spheroid cultures, we identified shared expression programs that relate to intestinal lineages and revealed metabolic signatures that are linked to cancer cell differentiation. In addition, we validated and complemented sequencing results by quantitative microscopy using live-dyes and multiplexed RNA fluorescence *in situ* hybridization, thereby revealing metabolic compartmentalization and potential cell-cell interactions.

Taken together, we believe that our approaches provide a framework for translational research to dissect heterogeneous transcriptional programs in 3D cell culture systems which will pave the way for a deeper understanding of functional tumor cell heterogeneity.

Zusammenfassung

Dreidimensionale (3D) *in vitro* Zellkultursysteme haben maßgeblich die Modellierung zellulärer Prozesse verbessert, indem physiologische Eigenschaften und strukturelle Merkmale von *in vivo* Geweben besser reflektiert werden. Darauf basierend entwickeln sich nun vermehrt repräsentative patientenabgeleitete 3D-Kultursysteme als verbesserte präklinische Tumormodelle, um individualisierte Therapieansätze zu unterstützen. Neben den zusätzlichen Fortschritten, die durch molekulare und pathologische Analysen hinsichtlich personalisierter Behandlungen erzielt wurden, verbleibt sowohl in primären Tumoren als auch in *in vitro* Zellkultur Systemen das begrenzte Verständnis der funktionellen Tumorzellheterogenität, was zunehmend als Schlüsselfaktor für Tumorprogression und Behandlungsresistenz erkannt wird. Neueste technologische Fortschritte basierend auf Next-Generation-Sequencing (NGS) ermöglichen nun Genom-weite Genexpressions-analysen in Proben mit geringem RNA-Gehalt und sogar Einzelzellen (scRNA-seq). Somit wurde die Grundlage geschaffen, sowohl zelluläre Subtypen aufzudecken als auch Gene zu identifizieren, die spezifisch Zellzustandsübergänge antreiben. Diese Ansätze erfordern jedoch im Allgemeinen die Dissoziation von Geweben in Einzelzellen, was folglich zum Informationsverlust multizellulärer Zusammenhänge führt. Daher wird grundsätzlich eine direkte oder indirekte Kombination von RNA-Sequenzierung und Mikroskopie für Einzelzellanalysen benötigt, um die Assoziation zwischen komplexen zellulären Phänotypen und ihren zugrunde liegenden genetischen Programmen genau zu verstehen.

In dieser Arbeit präsentiere ich zwei komplementäre Strategien basierend auf der Kombination von Mikroskopie und NGS, um die Heterogenität von Tumorzellen in 3D-Zellkultursystemen zu analysieren. Zunächst werde ich die Entwicklung und Anwendung der neuen Methode "pheno-seq" beschreiben, in welcher Hochdurchsatz-Bildgebung und RNA-Sequenzierung klonaler Tumor-Sphäroide direkt kombiniert wird. Durch diesen Ansatz konnten wir charakteristische Genexpressionssignaturen in 3D-Modellen von Brust- und Dickdarmkrebs nachweisen, die mit heterogenem invasivem und proliferativem Verhalten assoziiert sind. Zudem konnten wir Transkriptionsregulatoren identifizierten, die mit Hilfe von scRNA-Seq nicht identifiziert werden konnten, aus visuellen Phänotypen und assoziierten Genexpressionssignaturen Inhibitorantworten vorhersagen und regulatorische Einzelzellzustände errechnen. Zweitens haben wir durch Anwendung von scRNA-seq auf 12 Patienten-abgeleitete Kolorektalkrebs-Sphäroidkulturen gemeinsame Expressions-programme identifiziert, die sich auf intestinale Subtypen beziehen und konnten

metabolische Signaturen aufzeigen, die mit der Krebszelldifferenzierung in Verbindung stehen. Darüber hinaus haben wir Sequenzierungsergebnisse durch quantitative Mikroskopie unter Verwendung von Fluoreszenzfarbstoffen und multiplexed RNA-Fluoreszenz-in-situ-Hybridisierung (FISH) validiert und ergänzt, wodurch eine metabolische Kompartimentierung und mögliche Zell-Zell-Wechselwirkungen aufgedeckt werden konnte.
Wir sind davon überzeugt, dass unser Rahmenkonzept zur Analyse der zellulären Heterogenität in 3D-Zellkultursystemen mithilfe von Kombinationen aus NGS und Mikroskopie von hohem Wert für die translationale Forschung ist und den Weg für ein detaillierteres Verständnis der intratumoralen Heterogenität ebnen wird.

Contents

1 Introduction

1.1 Three-dimensional *in vitro* cell culture systems

1.1.1 Cellular *in vitro* models: Controlling environment, space and time

Understanding biological systems in detail often requires modelling of cellular processes outside of their natural environment, generally defined as '*in vitro*' cell culture. This term today mainly refers to human or animal cells that are cultivated, expanded and passaged in defined media and experimental setups. Although real physiological conditions are always preferred, visual examination by microscopy, molecular analyses, (epi)genetic manipulation and drug perturbation approaches require reproducible, flexible and cost-effective experimental setups that are often not available for living multicellular eukaryotes. This holds especially true for human tumor samples for which *in vitro* culture confers the ability to observe dynamic cellular processes over time in multiple replicates.

In the last 50 years, the application of *in vitro* cell culture systems has expanded in nearly every field of biological and biomedical research, including developmental and stem cell biology, disease modeling, drug discovery and regenerative medicine[1]. Until recently, most widely used experimental setups include classical two-dimensional (2D) cell culture systems that are mainly based on immortalized or cancer cell lines that are grown on solid, impermeable surfaces. Despite their obvious limitations in reflecting three-dimensional (3D) tissue physiology, 2D cell culture systems have greatly contributed to the understanding of basic principles in biology. Furthermore, many of these are still widely used for methodological development and proof-of-concept studies due to their convenience in culture and maintenance[2]. However, several alternative and more physiologically relevant human cell culture models were developed in the last decades which are progressively replacing standard 2D *in vitro* methods.

1.1.2 From 2D to 3D: Modeling tissues in three dimensions

A major limitation of adherent 2D monolayer cultures is their lack of physiologic tissue geometry and architecture that fails to reflect cellular and microenvironmental interactions. Cell-cell and cell-extracellular matrix (ECM) contacts as well as bio-mechanical cues mediate specific transcriptional programs and signaling cascades that are required for stem cell maintenance and functional differentiation[3–6]. Thus, environmental stimuli and cellular heterogeneity that characterize primary tissues are severely limited in 2D culture systems, which most probably explains why pre-clinical drug-screening in 2D does often not reflect *in vivo* outcomes[7,8].

To overcome these limitations, the first efforts to culture cells in three dimensions have been made already 70 years ago by isolating and maintaining primary mammary glands from mice[9]. Major landmark publications that provided the basis for today´s research include the discovery and characterization of main components of the ECM[10,11], as well as the isolation of laminin-rich matrix hydrogels from chondrosarcomas[12], now widely used as the basement membrane surrogate (Matrigel). In the late 80´s and 90´s, pioneering work by Mina Bissell and co-workers revealed principles and mechanisms of how ECM components regulate cellular morphologies and differentiation in the mammary gland[3,7,13,14]. They also developed standard Matrigel-based assays to study mammary morphogenesis that revealed phenotypic differences between healthy and tumor cells[15,16]. Alternatively, also a variety of spherical ECM-independent floating 3D culture models were established to model physiologic characteristics of tissues and tumors or to enrich for (cancer) stem cells[17].

In 2007, two milestone studies transformed the fields of adult stem cell research and tissue modeling in 3D. First, Takahashi et at. generated induced pluripotent stem cells (iPS) from dermal fibroblasts that can be differentiated into all three germ layers *in vitro*[18]. Six years later, this tool was used to generate cerebral organoids based on skin fibroblasts from a patient with microcephaly[19]. Second, Barker et al. identified intestinal stem cells at the base of the crypts by the marker gene LGR5 (Leucin-rich repeat–containing G protein–coupled receptor 5)[20]. Based on this finding, they adapted Matrigel-based protocols from mammary gland literature[15] to generate clonal self-organizing 3D organoids with crypt-villus architecture and all differentiated cell types of the intestinal epithelium[21,22]. Furthermore, they designed a serum-free culture system that mimics the *in vivo* stem cell niche enabling long-term maintenance of intestinal organoids. Their results indicate no inherent restriction of the replicative potential of adult stem cells *in vitro*[23] and their methodological strategy could be successfully transferred to other tissues, including the stomach[24], pancreas[25] and liver[26], respectively.

1.1.3 3D cell culture system in translational cancer research

Our understanding of the origin and progression of cancer has significantly increased during the last decades. Despite parallel advancements in treating many types of cancer, it remains a major health problem worldwide[27]. One limiting factor is the high variability of cellular drug responses between (intertumor heterogeneity) and within single patients (intratumor heterogeneity) that severely complicates therapy decisions (detailed description in section 1.1)[28]. Thus, the development of new personalized therapies appears as a key strategy to effectively treat cancer that consequently requires physiologically relevant human cancer

models. However, standard 2D *in vitro* culture systems based on cancer cell lines have been the most widely used models for drug screening and a high proportion of drugs that perform well in these setups fail in clinical trials[29]. Alternatively, animal cancer models have greatly contributed to understand basic disease mechanisms, the high costs, time consuming generation and low throughput limit their broad applicability.

More recently developed 3D *in vitro* culture methods (section 1.1.2) do not only provide a more physiologic basis for research on adult stem cells and tissue homeostasis than classical 2D models, but also for translational cancer research. The general approach involves the isolation of tumor and/or normal cells from a patient as well as the subsequent 3D culture, passaging and cryopreservation. Optimally, the success rate in establishing cultures from different patients should be high (>50%) and cells should remain genetically and phenotypically stable over time. Established cultures can then serve as patient-specific 'Avatars' that can be used in combination with various other molecular tools and profiling methods to understand disease mechanisms and cellular heterogeneity (Figure 1.1).

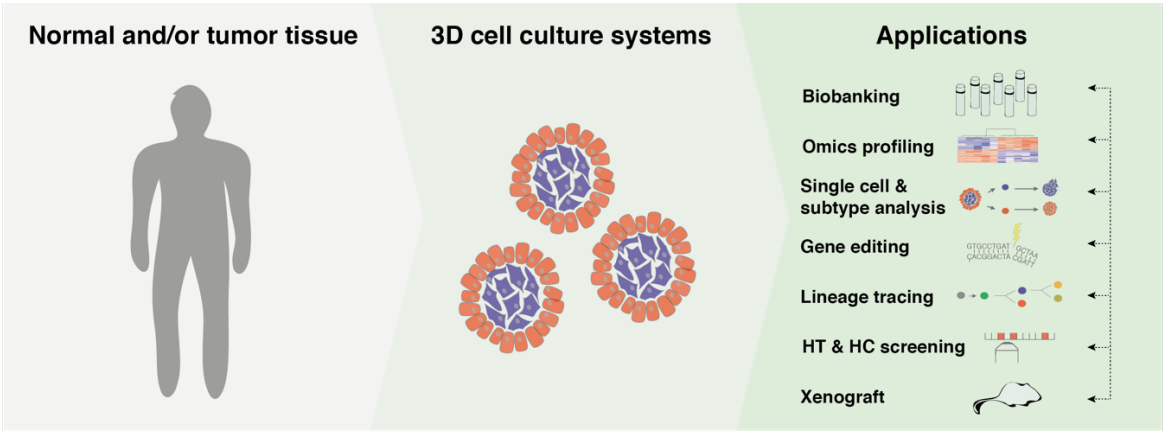


Figure 1.1 | 3D cell cultures systems in translational research. Schematic overview of possible applications for patient-derived 3D cell culture systems. Omics profiling involves genomic, transcriptomic, epigenomic, proteomic and metabolomic profiling. Many of the listed methodological strategies can be applied simultaneously, in pairwise or multiple combinations (dotted arrows). (HT: high-throughput, HC: high-content)

1.1.3.1 Spherical tumor models

Free-floating tumor spheroids cultured in serum-free medium supplemented with growth factors represent are one of the two most widely used patient-derived 3D *in vitro* models (also called tumorospheres or according to their origin colospheres, mammospheres or neurospheres, respectively). First described in 2003, the culture method has been initially developed for the expansion of cancer stem cells (CSCs)[30]. While non-malignant cells are

depleted due to inhibition of adherence and subsequent anoikis, it is generally assumed that single CSCs are able to generate continuously growing spheroids whereas more-differentiated cells have a limited proliferative capacity. However, this hypothesis still needs to be proven in detail. Although other types of floating spheroid models exist, this kind of 3D model has been highly valuable for the study of CSCs[17]. Floating tumor spheroids are often used in combinations with serial xenotransplantation in immunocompromised mice in order to assess the presence of tumorigenic CSCs[31] (see section 1.3.1.2 for more details on CSCs). However, little is still known about the biology of free-floating spheroids, including presence and proportional composition of subtypes and the grade of differentiation across patient cultures.

1.1.3.2 Patient-derived cancer organoids

Stem cell-derived organoid cultures that are based on the same protocols as those for non-malignant cells are probably the most widely used 3D culture systems for translational cancer research[32]. Although the methods are more cost and time intensive than those for free-floating spheroids, their physiologic relevance might be increased due to the presence of ECM, their highly defined medium composition and their high efficiency in establishing cultures from different patients. In order to deplete for non-malignant cells, the main strategy involves culture under selective growth conditions. For example, tumor cells with mutations in the epidermal growth factor receptor (EGFR) signaling pathway can be selected by EGF withdrawal[33]. Currently, many biobanks are generated from large collections of patient-derived tumor and matching healthy organoids. These resources can then be used for personalized medicine, including drug screening and NGS, as well as for the development of tumor specific therapies[8,34,35]. Moreover, healthy organoids have been used for disease modeling by targeted gene editing of the most commonly mutated CRC genes, thereby showing that organoid growth becomes independent of stem cell niche factors upon successive incorporation of mutations[36]. Furthermore, patient-derived CRC organoids have been used for lineage tracing of putative CSCs in combination with xenotransplantation[37]. However, relatively little is known about the biology and composition of tumor organoids, similar to floating spheroid cultures. Taken together, 3D cell culture systems significantly improve the physiologic complexity compared to classical 2D methods and at the same time maintain high experimental flexibility. Thus, they represent attractive tools for translational cancer research and personalized medicine.

1.2 Single cell analysis

1.2.1 Cellular phenotyping and *in situ* analysis by microscopy

Cells are the constituents of life in all organisms. For example, the human body consists of approximately 3.72 x 10^{13} cells with distinct functional roles that define human physiology and when perturbed result in diseased states like cancer[38]. Robert Hooke was the first to describe 'cells' in plants in 1665 with the microscope invented by Antoni van Leeuwenhoek. However, it took almost two centuries (1838-1855) until the 'Cell Theory' was formulated, in which Rudolf Virchow and others stated that (i) all organisms consist of one or more cells; (ii) that cells are the basic unit of life; (iii) and cells derive from pre-existing cells[39]. Following this guideline, biological research then aimed to classify and characterize cellular subtypes based on various properties that were, until recently, mainly detected by light microscopy.

For over 150 years, visual observation of contextual cellular phenotypes *in situ* represents a common strategy. In order to overcome optical restrictions in imaging whole tissues, physical sectioning (histology) or *in vitro* cell culture are the most widely used preparation techniques to analyze cellular heterogeneity, although intravital microscopy[40] and optical clearing[41] are potential alternatives for deep tissue imaging. Whereas early work in staining histological sections mainly involved the analysis of general tissue architecture and cellular morphologies, the advent of immunohistochemistry (IHC)[42], monoclonal antibodies[43] and *in situ* hybridization techniques[44] enabled researchers to more precisely distinguish subtypes based on molecular markers. With time, it became more and more evident that different molecular profiles usually define distinct functionalities even if cells were morphologically indistinguishable from one another. This step was the onset to understand biology at the systems-level as molecular characteristics are the direct consequence of underlying genetic programs.

The invention of the fluorescence microscope[45] alongside the development of additional molecular staining tools based on fluorescently-labelled antibodies[46], fluorescent dyes[47] and fluorescent proteins[48] where the key technological advances in molecular imaging. The general principle of enhancing the contrast by using fluorophores that emit light at different wavelengths to the excitation wavelength as well as the usage of dichroic mirrors[49] shaped the basic design of microscopes that are now standard in most biological laboratories. Several innovative illumination and detection strategies have been developed, including confocal[50], two-photon[51] and light sheet fluorescence microscopy[52], that enable imaging at higher spatial-temporal resolution. Image-based profiling of cells has now evolved towards a quantitative science[53] and recent technological advances in microscopy automation and data

analysis enable high-throughput phenotyping to quantitatively characterize cellular heterogeneity[54].

RNA fluorescence in-situ hybridization (RNA-FISH)[55,56] has developed into the method of choice for quantitative measurements of transcript abundance *in situ*. The general principle relies on the design of fluorophore-labelled oligonucleotide probes that are complementary to the RNA of interest. One key advantage compared of RNA-FISH over immunofluorescence (IF) is its high specificity and robustness because the binding of the designed probe is sequence dependent and not limited by the quality of an antibody. Thus, a broad range of hybridization probes can be used without significant differences in signal intensities. In addition, more recently developed methods for single molecule (sm)RNA-FISH[57] enable the detection of absolute mRNA copy numbers even if genes are expressed at very low levels.

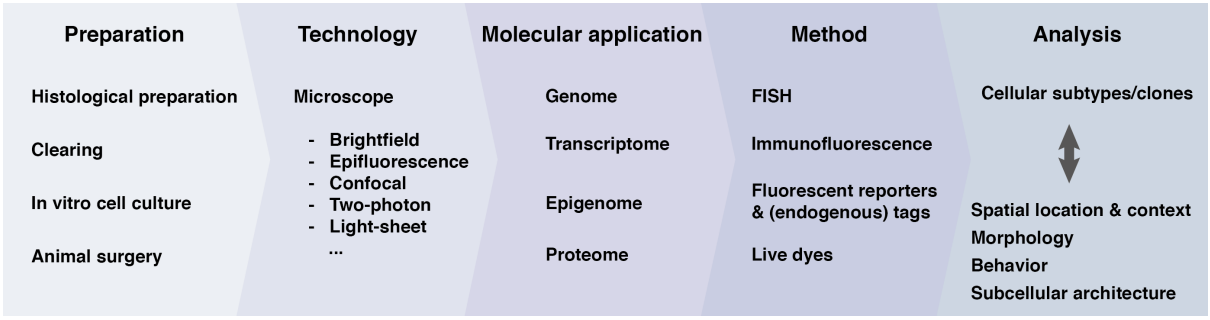


Figure 1.2 | *In situ* single cell analysis by microscopy. Summary of diverse imaging-based strategies for molecular single cell analysis *in situ* that generally require pre-selection of a limited number of defined markers but provide information of both cellular phenotypes and subtype-specific molecular features at the same time.

Despite the ability to detect molecular features and visual cellular *in situ* phenotypes in parallel (Figure 1.2), light microscopy-based methods are inherently limited by the number of molecular measurements that can be obtained from a single cell. Specifically, all these methods share the fundamental bottleneck in multiplexing due to the limited number of fluorescent probes with spectrally distinct fluorophores. To overcome this limitation, new methods for highly multiplexed fluorescence *in situ* hybridization (FISH)[58] and protein staining[59] have been developed that enable parallel detection of 10´s to 100´s of different molecular features per cell. However, these methods require highly complex experimental setups and pre-selection of transcript-specific probes or protein-specific antibodies. Thus, alternative methods are needed to enable the unbiased detection of molecular features in single cells in order to more systematically define cellular subtypes.

1.2.2 Single cell sequencing

1.2.2.1 Towards whole transcriptome analysis by RNA-sequencing

A single somatic cell usually contains two copies of its DNA, approximately 50,000 – 300,000 mRNA molecules[60] and millions of proteins[61]. Optimally, one would like to acquire an unbiased and system-wide view of all molecular components (genomic, epigenomic, transcriptomic, proteomic and metabolomic) in single cells to fully understand the causal relationships between genetic variation, regulatory mechanisms and phenotypic outcomes in heterogeneous populations. However, obtaining even one layer of information from minute amounts of molecules in single cells is technically challenging (DNA: approximately 6 pg, total RNA: 5-30 pg[62], protein: approximately 20-200 pg[63]).

Single cell transcriptomics, a general term for methods to quantitatively measure the abundance of RNAs, has evolved as a major strategy for multiplexed or unbiased measurement of gene expression. Technically, RNAs can be targeted specifically by sequence and/or reverse transcribed to more stable and easily amplifiable complementary (c)DNA. As the transcriptome represents the first output layer of gene expression, it can serve as fingerprint to reveal subtype identity and associated genetic markers, lineage relationships as well as underlying regulatory networks (see section 1.2.2.3). Initial methods for quantitative single cell transcriptomics include smRNA-FISH (see section 1.2.1) and single cell quantitative PCR (sc-qPCR)[64,65], of which the latter involves the conversion of RNA to cDNA by reverse transcription (RT) followed by cDNA amplification and quantification. Until today, both approaches still represent the gold-standard for targeted analysis of transcripts in single cells although they require pre-selection of transcript-specific probes or primers, respectively.

In contrast, unbiased whole transcriptome analysis was first restricted to cellular bulk measurements due to the required input amount of RNA. The first step towards this goal was made by the development of microarrays[66], a hybridization-based approach that involves the incubation of fluorescently labelled cDNA with custom-made arrays of complementary DNA sequence. However, a key advancement towards unbiased analysis of whole transcriptomes at base pair resolution was made by the development of next generation sequencing (NGS) platforms[67], of which the bridge amplification and reversible termination technology (Illumina) evolved as a standard worldwide. The methodological principle for RNA-sequencing (RNA-seq) by Illumina NGS relies on the ligation of adaptor sequences to the ends of fragmented cDNA that can bind to covalently attached primers on a glass flow cell. Upon binding, single cDNA molecules are then amplified by bridge amplification to form clusters of clonal sequences. These populations of identical templates (usually 100 – 500 million clusters)

then undergo the actual sequencing reaction with reversible terminator chemistry: The major steps involve (i) the addition and incorporation of all four nucleotides each labelled with a different dye, (ii) washing to remove unbound nucleotides, (iii) fluorescent readout by imaging, (iv) cleavage reaction to remove dye and terminating group and (v) washing. By successive rounds of base incorporation, washing and imaging followed by image analysis, single nucleotide signals derived from a cluster of clonally amplified sequences are assembled to short reads (typically 30-250 base pairs). These can then be aligned to a reference genome to generate a nucleobase-resolution gene expression profile[68].

1.2.2.2 Methodological strategies for single cell RNA-sequencing

Despite the major improvements in transcriptomic analyses enabled by RNA-seq, initial protocols based on bulk measurements share the fundamental limitation of averaging signals from individual cells together. Consequently, crucial information of subtype specific expression is lost that can lead to misinterpreting data[69]. The first time RNA-seq was adapted for the analysis of single cells was in 2009 (scRNA-seq)[70] - only 2 years after this method was applied to bulk populations of cells. Generally, scRNA-seq requires different sample preparation workflows compared to above described *in situ* methods (section 1.2.1). Primarily, cells need to be first dissociated from tissues, followed by isolation of single cells in suspension and subsequent molecular profiling (Figure 1.3). Although these approaches lead to a loss of information regarding tissue context and cellular morphology, the placement of individual cells in defined and independent reaction volumes facilitates higher throughput and precision for transcript processing and measurement.

Capturing single cells with high efficiency and at high throughput is one of the key challenges in scRNA-seq workflows. Whereas cells have been picked manually by micromanipulation in early studies[70], technological advances now enable the isolation of hundreds to thousands of cells in a single experiment[71,72]. One of the earliest techniques to isolate and analyze cells in suspension is fluorescence activated cell sorting (FACS)[73,74], which not only enables fluorescent measurements, but also the separation of cells based on their molecular properties. This approach turned out to be a powerful way to understand the hematopoietic and immune systems[75], especially because the isolation of these cell types usually don´t require dissociation. Although multiplexed molecular detection in FACS analyses faces the same optical restrictions as other light microscopy methods, isolation of tens to hundreds of cells by flow cytometry still represents one of the most widely used methods for single cell capture in scRNA-seq workflows[76].

Automation as well as miniaturization of reaction volumes are key improvements to more recently developed scRNA-seq technologies. Aside from the fact that the reduction to nanoliter scale reaction volumes dramatically decreases the reagent cost, it has also been shown that lower reaction volumes have a positive influence on accuracy and sensitivity, most probably due to a higher effective concentration of the reactants[77,78]. Typically, this has been achieved by the use of microfluidics[79–81] or well-based[82,83] platforms (further description below).

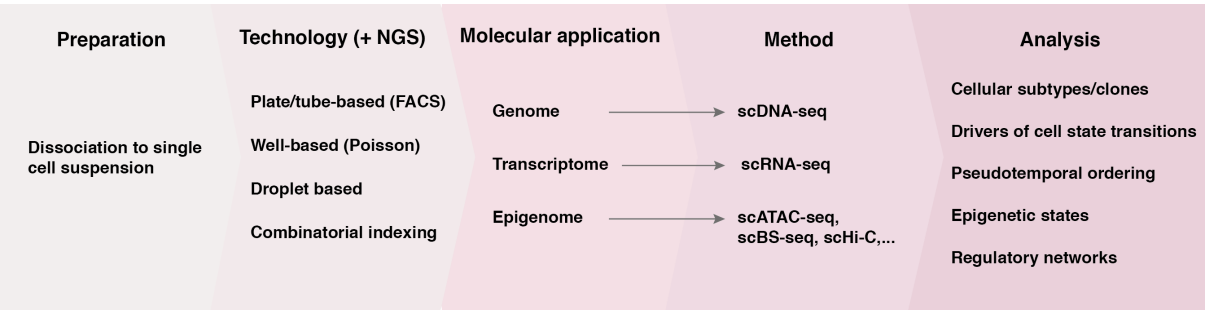


Figure 1.3 I Single cell analysis by next-generation sequencing. Overview of dissociation-dependent strategies for single cell analysis based on next-generation sequencing for unbiased identification of subtype-specific molecular features. Plate/tube-based assays generally require fluorescence-activated cell sorting (FACS). In Well-based setups (e.g. SeqWell[82] or iCELL8[84]) cells are isolated by limiting dilution, thereby assuming Poisson-based distribution per well.

Most of the current scRNA-seq protocols capture polyadenylated RNA species by using poly(T) primers to initiate reverse transcription (RT). The efficiency of this key step varies between different scRNA-seq methods. It is estimated that approximately 5-40% of transcripts are reverse transcribed, resulting in high technical noise, especially for lowly expressed genes[78]. Second-strand synthesis is typically achieved by template switching at the 5´-end of the transcript[85] to generate full-length cDNA. Next, the minute amounts of cDNA need to be amplified either by PCR or by *in vitro* transcription to acquire enough material for library preparation and NGS.

The introduction of barcodes during generation of scRNA-seq libraries is inevitable for highly parallel single cell processing and multiplexed sequencing. Depending on the strategy for NGS library preparation, two kinds of library structure and transcript data have been mainly generated during the last years. First, SmartSeq[85] and SmartSeq2[86] protocols enable the generation of full-length libraries to additionally facilitate the analysis of alternative splicing and allele-specific expression. However, since index barcodes for multiplexing are introduced late during sequencing adapter ligation after cDNA amplification, full-length

protocols are limited in parallel processing or require sophisticated robotic workflows to achieve high-throughput. Alternatively, cell specific barcodes can be introduced already during RT with the Poly(T) primer. This strategy enables a much higher throughput, as libraries can be pooled at early time points which reduces reagent cost and hands on time. However, sequencing of these libraries is restricted to the counting of 3´- or 5´-ends.

In general, most studies employing full-length workflows use FACS to isolate and process cells in 96 or 384 well plates[87–89]. Alternatively, one of the first commercially available and automated scRNA-seq systems is the laminar-flow microfluidics system Fluidigm C1, capable of generating up to 96 full-length libraries in a single experiment[79]. More recently, droplet-based microfluidic instruments for 3'-end counting was developed to capture thousands of single cells in individual partitions with barcoded beads[80,81], thereby increasing the throughput by at least one order of magnitude per assay. Most of the droplet-based platforms are already commercially available (10x Genomics Chromium, 1CellBio InDrop, Dolomite μEncapsulator). Alternatively, methods based on distribution of cells into micro- or nanowells, either by manual pipetting (SeqWell[90]) or solenoid valve dispensing (TakaraBio iCELL8[83]), also achieve profiling of >1000 cells per experiment. Whereas SeqWell uses barcoded beads similar to droplet-based methods, barcoded Poly(T) primers are pre-printed in nanowells of the iCELL8 system. In general, cell capture in well-based setups relies on limiting dilution, assuming Poisson-based distribution per well. A major advantage of the iCELL8 system is an integrated imaging step that enables automated evaluation and selection of cells for sequencing based on their visual properties. Lastly, 'split and pool' barcoding methods based on combinatorial indexing of single cells are the newest generation of scRNA-seq methods[91]. Notably, these methods don´t require complex experimental setups and enable transcriptional profiling of $>10^4$ cells per experiment.

In addition, other methods for unbiased genomic[92], epigenomic[93–95] or even multimodal[96–100] molecular profiling of single cells have been developed recently. However, a detailed description of these methods is not provided here due to the lack of relevance for the results of this work.

1.2.2.3 Computational analysis of scRNA-seq data

Although computational analysis workflows for bulk and single cell transcriptomic NGS data share many similarities including read pre-processing, alignment and generation of read counts, scRNA-seq data has unique characteristics and its analysis is associated with specific statistical challenges[101]. First, the sample numbers of published single cell datasets increased dramatically during the last years, ranging from hundreds to $>10^5$ cells and thus

strongly exceed classical bulk RNA-seq datasets in size. Moreover, scRNA-seq data is characterized by substantial technical noise that can be introduced by dissociation[102], inefficient cell lysis, low RT efficiency[78], cDNA amplification (up to 1 million fold) and sequencing. A major consequence of technical noise due to the low starting material are gene 'dropouts'. It describes the case in which a transcript is not detected because it has not been captured or amplified although it is present in the profiled single cell. This mainly affects lowly and moderately expressed genes and leads to zero-inflated expression matrices[103]. Another major challenge is the presence of biological noise (e.g. oscillating processes like the cell cycle) that strongly contribute to gene expression heterogeneity and can therefore complicate the identification of cellular subtypes[104]. Thus, scRNA-seq analysis requires careful examination in order to understand cellular heterogeneity and to avoid misinterpretations.

After the generation of read counts, typical scRNA-seq workflows consist of the following steps: (i) removal of low-quality libraries that are usually characterized by a low amount of total reads and by high numbers of mitochondrial reads[105,106]; (ii) count normalization to correct for differences in library complexity or to account for other technical or biological confounders[107]; (iii) 'feature selection' to remove uninformative (typically low-expressed) genes[108–110]; (iv) dimensionality reduction which is usually achieved by principal component analysis (PCA)[104,111] or T-distributed stochastic neighbor embedding (tSNE)[112]; (v) computation of pairwise cell-cell distances and clustering (e.g. k-means) to identify cellular subpopulations[113]; (vi) and characterization of subpopulations (e.g. by identification of subtype markers by differential expression analysis[109,114]). The last two steps of grouping and characterizing cells represent the most popular applications in scRNA-seq analysis. Biologically, identified populations can be distinct cell types (e.g. epithelial vs. stromal fibroblasts in the intestine), or they correspond to different states of the same cell type (e.g. activated vs. non-activated T-cells).

During the last years, the number of computational tools developed for the analysis of scRNA-seq data are increasing, but no gold-standard has evolved, yet. However, several toolkits have been developed to enable the streamlined analysis and exploration of scRNA-seq data including abovementioned steps and their multiple combinations[105,115]. More recent tools focus on correlated gene expression for the 'feature selection' step by testing annotated and *de novo* identified gene sets for coordinated expression variability across cells[116,117]. Aside from utilizing prior knowledge, this strategy facilitates the identification of overlapping aspects of gene expression heterogeneity and the removal of unwanted confounders.

Furthermore, more specific scRNA-seq analysis tools have been developed that go beyond subtype identification, including the pseudo-temporal ordering of cells and the identification of drivers of cell state transitions[118,119], reconstruction of transcriptional networks[120,121], analysis of allelic gene expression[122] and detection of alternative splicing[123]. Although scRNA-seq is a relatively young technology that still requires improvements for both experimental and computational workflows, it provides a promising basis to understand cellular heterogeneity in a systematic and unbiased way. However, scRNA-seq methods should be directly or indirectly combined with imaging approaches *in situ* in order to understand contextual phenotypes, as spatial information and cellular morphologies are lost during sample preparation.

1.2.3 Combinations of NGS and microscopy for the analysis of cellular heterogeneity *in situ*

Cellular fate and behavior are mainly governed by the spatial location, which directly influences its gene expression by external molecular signals and interactions. In addition, contextual morphological features are highly informative for cellular function that is a direct consequence of its underlying regulatory gene expression network. Understanding these relationships represents a main focus of developmental and translational research but requires sophisticated genomic tools to dissect these processes. Although scRNA-seq alone can in principle be used to predict cellular interdependencies based on anticipated receptor-ligand interactions[124], this approach requires prior-knowledge and is therefore limited for the discovery of new interactions. Moreover, cellular morphologies and spatial contexts are lost upon dissociation and therefore cannot be correlated with gene expression profiles. Thus, combining transcriptome-wide gene expression profiling with visualization of morphological features and physical cell-cell interactions *in situ* is required to fully understand complex tissue biology.

When subpopulations have been identified with scRNA-seq, RNA-FISH, IF or fluorescent reporter genes are commonly used to link subtypes to their spatial location or morphology in an independent experiment in order to compensate for the lost cellular context. This strategy is most often used for qualitative validation of scRNA-seq data[22,125,126], but several studies also quantitatively integrate both datasets, for example by mapping cell types to specific 3D coordinates in model organisms[127,128]. Although this indirect combination has contributed significantly to understand complex tissue biology, imaging-based methods for molecular *in situ* profiling are usually limited to a handful of pre-selected markers and are therefore limited in the number of subtypes that can be mapped in parallel. Moreover, single cell

transcriptomes in complex biological systems are not solely characterized by discrete cell types, but also by oscillating or transient cell states (e.g. cell cycle[104] or metabolism) as well as by continuous differentiation processes[129] and transcript gradients[128]. Thus, mapping a few cell types by single selected markers will only give an incomplete picture of contextual cellular phenotypes and underlying gene expression. As mentioned previously, highly multiplexed RNA-FISH approaches (see section 1.2.1)[58,130] would represent powerful tools to dissect gene expression *in situ,* especially in combination with scRNA-seq. However, they require highly complex and specialized experimental setups which limits their wide applicability.

Method	Laser capture microdissection	Spatial transcriptomics	FISSEQ/StarMap
Workflow	Histological preparation & imaging ▼ Laser dissection and capture of manually defined region ▼ RNA extraction, RT and cDNA amplification ▼ Library preparation & Sequencing	Histological preparation & imaging ▼ Placement of tissue slice on Poly-T primer array with regional barcodes ▼ Permeabilization, mRNA capture and cDNA synthesis/amplification on array ▼ Library preparation & Sequencing	Culture/Histological preparation ▼ In situ reverse transcription and crosslinking ▼ cDNA rolling-circle amplification ▼ Consecutive rounds of (re)hybridization and imaging
Detected genes per feature	~10^2 - 10^3	~10^3 - 10^4	~10^1 - 10^3
Cellular resolution	>1 cell	10-20 cells	single cell
Workflow complexity	moderate	low	very high
Throughput	low	high	average
Application	3D cell culture & tissue sections	Tissue sections	2D cell culture & tissue sections

Figure 1.4 I Hybrids of imaging and sequencing. Overview of existing methods that directly combine microscopy and RNA sequencing for unbiased identification of gene expression and associated contextual cellular phenotypes at the same time.

To overcome these limitations, 'hybrid' approaches have been developed that directly combine microscopy and unbiased gene expression profiling (Figure 1.4). For example, laser capture microdissection (LCM) describes a method that enables the isolation of cells from histological slices by laser cutting, which has already been combined with low input gene expression profiling in archived frozen tissue[131] and 3D cell culture systems[132]. Although this strategy maintains the spatial information of isolated cells, key limitations are the low throughput and associated low sample quality that usually requires gene expression

profiling of multiple cells in order to acquire enough material for cDNA generation. Alternatively, recent studies have demonstrated the direct combination of histology and RNA-seq on primary samples to spatially and morphologically resolve intratumor heterogeneity (spatial transcriptomics)[133,134]. This method relies on a Poly(T) primer array with regional barcodes that are distributed in 100 μm spots, thereby reaching a transcriptomic cellular resolution of 10-30 cells. However, this method is limited in cellular resolution and is not suited for 3D cell cultures systems. Finally*, in situ* RNA-sequencing describes methods that sequence transcripts directly in tissue sections[135] or in 2D cultured cells[136]. This is achieved by hydrogels that are used for crosslinking of cellular components, followed by RT, cDNA amplification and consecutive rounds of hybridization and imaging. Although these methods most likely represent the future of whole tissue imaging, they require highly elaborate sample preparation and imaging technology similar to multiplexed RNA-FISH. Additionally, available *in situ* sequencing publications only represent proof-of-concept studies that have not evaluated technical biases and restrictions in detail.

Taken together, the indirect combination of scRNA-seq and imaging with a limited set of pre-selected markers represents a straightforward strategy but is limited in resolution for *in situ* analysis. Alternatively, new methods for a direct combination of microscopy and unbiased gene expression profiling exist, but these technologies are only beginning to emerge or still suffer from technical limitations. In addition, there is no method yet to directly combine imaging and sequencing in 3D cell culture systems in a high-throughput manner.

1.3 Intratumor heterogeneity

1.3.1 Origin and consequences of intratumor heterogeneity

Tumor development is characterized by successive (epi)genetic alterations that activate oncogenes and inactivate tumor suppressors. These lead to the progressive acquisition of biological capabilities (generally described as the 'hallmarks' of cancer[137]) where cells evolve from normal to neoplastic, and finally to malignant states. However, the pattern and succession of mutational profiles and epigenetic changes as well as associated phenotypic outcomes do not only vary between different tumor entities and patients (intertumor heterogeneity), but also between cancer cells in a single tumor or patient[28] (intratumor heterogeneity). In addition, homotypic and heterotypic interactions of cancer, stromal and immune cells in the tumor microenvironment seem to have fundamental roles in cancer progression[138]. Thus, single cell analyses appear as key strategies to dissect cellular heterogeneity in tumors.

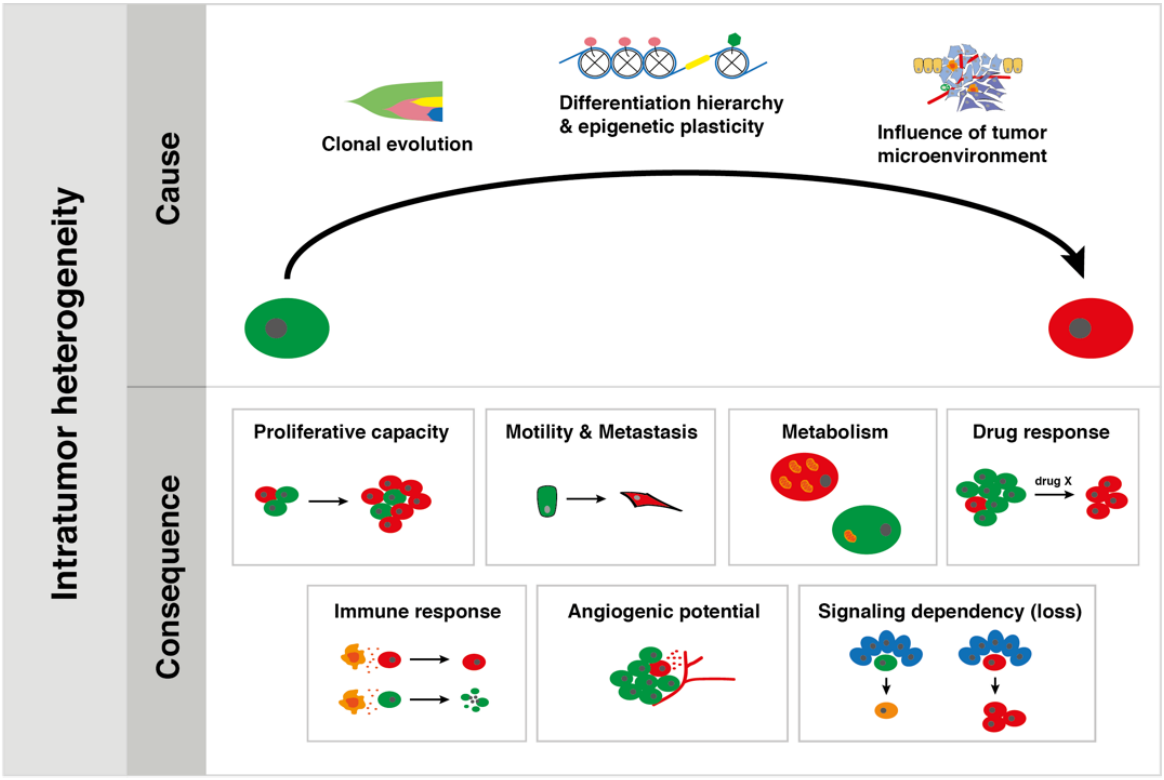


Figure 1.5 I Causes and consequences of intratumor heterogeneity. Upper: Schematic representation of factors influencing in tumor cell heterogeneity. Differences in cancer cell states can originate from genetically distinct subclones, from perturbed epigenetic regulation or developmental programs similar to those occurring during normal tissue homeostasis, and from regional differences of the tumor microenvironment. Lower: Functional consequences of heterogeneous cancer cell states affecting tumorigenicity including differences in long-term proliferative capacity, local tissue invasion and metastasis, metabolic preferences and dependencies, susceptibility and resistance to immune responses, angiogenic potential, loss of niche dependencies and drug resistance.

Large-scale studies based on bulk NGS have contributed significantly to understand interpatient heterogeneity[139,140], but these approaches, although a fundamental resource, are limited in revealing molecular variability between cells in individual tumors[141]. The remarkable progress in the development of single cell sequencing approaches (see section 1.2.2) has already started to transform research on intratumor heterogeneity by the ability to molecularly dissect tumors in an unbiased way[142]. In the following sections, I will summarize key findings of causes and consequences of intratumor heterogeneity (Figure 1.5), while focusing on the contribution of single cell analysis and sequencing. Finally, I will provide an overview of how *in vitro* cell culture systems have been used to study tumor cell heterogeneity.

1.3.1.1 Genetic alterations and tumor evolution

The acquisition of mutations and other genomic alterations is regarded as prerequisite for tumorigenic progression. The general theory implies that mutant cells gain selective advantages that enable clonal outgrowth. Thus, tumor evolution and progression can be described as consecutive clonal expansions that are mainly enabled by genomic instability - generating genetically heterogeneous cell populations[143]. This genetic heterogeneity in tumors also leads to therapeutically relevant phenotypic diversity, including drug resistance and clinical prognosis[144]. Initially, genetic intratumor heterogeneity and tumor evolution have been studied in different entities by multi-region DNA sequencing[145] or ultra-deep DNA sequencing[146]. Whereas sampling of multiple distinct regions is limited to the identification of clones that are actually spatially segregated, the inference of subclones by clustering of mutation frequencies from bulk data can be blurred by copy number variations (CNVs).

Technical advances in whole genome amplification from single nuclei for single cell DNA-seq (scDNA-seq) enabled a much more detailed view on tumor evolution, although these methods still suffer from technical errors and limited coverage uniformity[142]. Major insights have been acquired by the study of breast cancer by Nicholas Navin and co-workers, who could detect different clonal subpopulations that were missed by bulk approaches[147]. In a later study, they combined scDNA-seq and scRNA-seq to show that resistant clones were pre-existing before therapy and adapted their transcriptional profile in response to therapy[148]. Similar to the previously mentioned spatial transcriptomics technology, Casasent et al. developed topographic single cell sequencing (TSCS) for inference of copy number profiles from single tumor cells while preserving their spatial context[149]. By the combination of breast cancer histology, LCM and scDNA sequencing, the authors revealed a direct genomic

lineage of cells derived from ductal carcinoma *in situ* regions and those derived from invasive regions (further description on cancer cell invasion in section 1.3.1.4). In addition, they showed that genomic evolution occurs before invasion and that migration into adjacent tissues involved multiple clones, thereby demonstrating the power of combining molecular profiling and visual phenotyping.

Broad CNVs can be also detected from scRNA-seq profiles, either by averaging gene expression from larger chromosomal regions[87] or by using allelic gene expression information[150]. Interestingly, the latter study revealed transcriptionally distinct subpopulations in multiple myeloma patients that did not match the underlying subclonal structure, indicating for the presence of additional non-genetic alterations that drive tumor progression.

1.3.1.2 Epigenetic heterogeneity and the cancer stem cell model

Beside genetic differences between cancer cells, phenotypic and functional heterogeneity in single tumors also arise from epigenetic changes. These can be induced by cell-intrinsic (e.g., mutations, developmental programs) or cell-extrinsic (e.g., microenvironmental cues) mechanisms that affect chromatin and DNA methylation states of cancer cells[151]. Heterogeneous epigenetic states in tumors seem to share many similarities to those occurring during cellular differentiation in healthy tissues[152]. For example, self-renewal in the intestinal epithelium[153] and the hematopoietic system[154] is based on well described long-lived adult stem cells that give rise to multiple short-lived subtypes with specialized function. Similarly, the cancer stem cell (CSC) model states that tumors are also hierarchically organized, with tumorigenic CSCs that fuel tumor growth and differentiate into non-tumorigenic progeny. However, recent studies indicate for a more complex concept of CSC biology, including microenvironmental dependencies and cancer cell plasticity[155].

CSCs have been defined by their long-term proliferative potential *in vitro*, by their ability to seed new heterogeneous tumors upon serial transplantation in immunodeficient mice (also termed tumor-initiating cells) and that they share expression profiles with normal stem cells[156]. Furthermore, observations that CSCs can exhibit a slow-cycling phenotype or have the ability to switch to quiescent states might explain failed chemotherapies and drug resistance[157]. Recent scRNA-seq studies could support the CSC model by linking identified tumor cell subtypes to developmental transcriptional programs[87,89,158], but the underlying mechanisms during tumorigenic progression are poorly understood.

The intestinal epithelium most probably represents the best described stem cell compartment in both healthy and neoplastic tissues. Under normal conditions, $LGR5^+$ stem cells reside at the base of the small intestinal crypt and migrate upwards where they

differentiate into more specialized cell types, including absorptive enterocytes, mucus-producing goblet cells and hormone-secreting enteroendocrine cells, respectively[159]. One exception are terminally differentiated Paneth cells ('deep crypt secretory' [DCS] cells in the colon) that intermingle with LGR5$^+$ stem cells at the crypt base where they provide important niche-factors including epidermal growth factor (EGF) as well as WNT and Notch ligands[4,160]. WNT ligands are especially essential to maintain the undifferentiated state of intestinal stem cells. This is supported by the fact that most CRCs are characterized by genetic changes that constitutively activate WNT signaling[161], thereby inducing a general crypt progenitor phenotype in CRC cells[162]. However, gene expression of WNT pathway components exhibit a high variability in individual cells of single tumors despite sharing the same mutational signatures, indicating that microenvironmental factors also influence tumor cell gene expression and differentiation[163].

Key insights have been acquired by the deletion of the tumor suppressor and negative WNT regulator APC in combination with lineage tracing in a mouse model. In this study, Schepers et al. showed that, similar to healthy tissues, LGR5 marks a Paneth cell-framed subpopulation that fuels the growth of developing adenomas, suggesting a similar niche dependency and developmental hierarchy in intestinal neoplastic tissue[164]. More recently, three studies investigated the functional role of LGR5$^+$ CSCs in CRC by using engineered or patient-derived organoid cultures, xenotransplantation, LGR5 lineage tracing and cellular ablation to demonstrate the essential role of LGR5$^+$ cells for tumor growth and metastasis[37,165,166]. In sum, the studies revealed (i) tumor initiating cell activity of LGR5$^+$ cells in serial transplantations, (ii) the presence of quiescence LGR5$^+$ tumor cells, (iii) re-expression of LGR5 in differentiated cells upon ablation of LGR5$^+$ cells and (iv) different roles of LGR5$^+$ cells in primary tumors and metastases.

Albeit these studies provide strong arguments for the CSC model in CRC, several questions remain. For example, little is known about the cellular composition of CRC and no study has yet reported detailed information about heterogeneous tumor cell subtypes identified by scRNA-seq. Furthermore, niche dependencies and molecular mechanisms that drive tumor cell differentiation and plasticity remain largely unknown but could provide promising targets to eliminate CSCs.

1.3.1.3 Influence of the tumor microenvironment

Tumors are not composed of a homogeneous mass of proliferating cells, but are rather complex tissues with multiple interdependent malignant and non-malignant cell types and states[138,167]. Because of this high grade of heterogeneity, single cell approaches have

already proven highly valuable to dissect the composition of the 'tumor microenvionment' in an unbiased way[87,168,169].

In general, microenvironmental interactions that influence cancer progression can be distinguished in homotypic interactions between tumor cells, and heterotypic interactions between malignant cells and non-malignant cells, including cancer associated fibroblasts (CAFs), endothelial cells and immune cells[138]. Referring to the hierarchical organization of CRC described in the previous section, homotypic interactions are most likely similar to niche dependencies occurring in healthy tissues. Supporting evidence has been provided by studies of different lung cancer entities, in which the authors could identify niche-promoting tumor subpopulations that are defined by the expression of WNT[170] and Notch[171] signaling components. Moreover, by using patient-derived xenograft and genetic engineering, Ebinger et al. could identify a rare subpopulation of drug-resistant cells with stem-like characteristics in acute lymphoblastic leukemia, whose dormant phenotype is dependent on its *in vivo* niche[172]. However, the characteristics of the niche itself have not been described in further detail, which is the case for virtually all tumor entities to date.

In contrast, a growing number of studies using scRNA-seq revealed stromal and immune cell types and interactions in the tumor microenvironment, including immune cell phenotypes[173] and T-cell exhaustion[87], endothelial gene expression signatures[174] as well as subtypes of cancer associated fibroblasts[167]. As current 3D cell culture systems restricted and immature in modeling non-malignant components of the tumor microenvironment, this topic will not be covered here in further detail due to the lack of relevance for this work.

Besides cellular interactions in the microenvironment, ECM composition and variations in oxygen supply are additional factors that can influence tumor cell heterogeneity and cancer progression[137]. Whereas the influence of different components of the ECM is currently difficult to assess with single cell methods, the cellular response to limited oxygen supply is well defined. Similar to normal tissue, oxygen supply in tumors is provided by the vascular system. Although tumors are typically characterized by sustained proliferation of endothelial cells, and consequently by continuous development of new blood vessels[137], the rapid proliferation of cancer cells often leads to uneven and inefficient vascularization. The development of hypoxic regions is then associated with phenotypic changes, including reduced proliferation rates, downregulation of oxidative phosphorylation (OXPHOS) and upregulation of glycolytic metabolism[175] (for further description of tumor cell metabolism see section 1.3.1.5). Importantly, hypoxia has been linked to poor survival, which might be explained by therapy-resistant dormant states or increased metastatic potential of tumor cells[176,177].

The presence of hypoxic single cell gene expression signatures has been validated in glioblastoma by using scRNA-seq[168]. In this study, the authors showed that cells highly expressing hypoxic genes are in a non-proliferating state, thereby supporting previous notions of hypoxic phenotypes. However, this study could not provide any spatial information and therefore lacks evidence for the direct link between regional hypoxic niches and identified gene expression programs.

1.3.1.4 Cancer cell invasion and EMT

Most malignancies originate from epithelial tissues, generating so called carcinomas (e.g., including tumors of the colon, lung, pancreas, breast and liver). Whereas tumors in early stages keep epithelial characteristics, cells of late-stage carcinomas gain phenotypic features that enables them to invade local tissue and finally to metastasize to other organs[178]. In general, metastasis represents a key event during tumorigenic progression that causes approximately 90% of cancer deaths[179].

During the earliest stages of metastasis, cancer cells need to down-regulate genes that sustain epithelial phenotypes and up-regulate genes that mediate motility and local invasion. Later on, this switch needs to be reversed in order to colonize distant sites of the body. On the mechanistic level, this seems to be achieved by hijacking the developmental gene expression programs 'epithelial-mesenchymal transition' (EMT), and its counterpart mesenchymal-epithelial transition (MET) that normally occur during embryogenesis and wound healing[180]. For example, extensive studies in breast cancer models emphasize the link between EMT, tumorigenicity and metastasis, where the depletion of EMT transcription factor families such as Snail, Twist and Zeb1 strongly inhibited metastatic dissemination from primary tumors[181,182]. Aside from its role in tumor cell dissemination, the EMT program appears to influence additional cellular functions in tumor progression, including drug resistance[183], immunosuppression[184] and stemness[185].

It is not clear yet whether genetic or non-genetic mechanisms confer the ability of local tissue invasion and dissemination, although several studies indicate for a central role of the microenvironment as well[180]. Moreover, relatively little is known about the dynamic gene expression changes that occur during the switch from epithelial to mesenchymal behavior of cancer cells, however first single cell and spatial gene expression studies could shed light on this process. Other than the previously mentioned topographic scDNA-seq approach[149] (see section 1.3.1.1), spatial transcriptomics could reveal EMT related genes in histologically invasive regions of advanced breast cancer[133]. Furthermore, Sidharth et al. used scRNA-seq to identify tumor cells expressing EMT-related transcriptional programs ('partial-EMT') in

metastatic head and neck squamous cell carcinoma that localize to the leading edge of primary tumors[169]. They showed that partial-EMT gene expression programs are predictive for nodal metastasis and pathological grade. However, advanced 3D *in vitro* cell culture systems might enable a more detailed view on dynamic cellular behavior and associated gene expression changes during EMT.

1.3.1.5 Metabolic heterogeneity

Both normal and tumor cells rely on the ability to transform nutrients to energy, mostly in the form of adenosine triphosphate (ATP), and to building blocks of cellular components, including proteins, lipids, nucleic acids and complex carbohydrates. This happens in complex sequences of biochemical reactions that can be reprogrammed depending on the state and function of cells[186]. Hence, metabolic pathways are strongly inter-connected with gene expression programs that lead to highly heterogeneous metabolic states and dependencies in both normal and tumor tissues[187,188]. Furthermore, metabolic states are strongly influenced by genetic or epigenetic alterations that affect the expression metabolic enzymes during cancer progression[189].

Glucose represents the main source of energy, which is first transformed to pyruvate via glycolysis, and subsequently to carbon dioxide (CO_2) in the mitochondria via tricarboxylic (TCA) cycle and OXPHOS. Although glycolysis is much faster in providing ATP and cellular building blocks, mitochondrial respiration by OXPHOS is approximately 18-fold more efficient in generating ATP[186]. However, OXPHOS generates toxic by-products in form of reactive oxygen species (ROS) that can, at high levels, impair cellular functions. Metabolic states can be affected by regional differences in oxygen supply. The general and simplified notion implies that normal cells favor OXPHOS under aerobic conditions but shift towards glycolytic metabolism under anaerobic conditions. This phenomenon can be also linked to cellular differentiation as observed for hematopoietic stem cells (HSCs). Since HSCs reside in hypoxic niches[190], they mainly utilize glycolysis instead of OXPHOS[191] in order to maintain their quiescent state and to minimize oxidative stress by ROS. On the other hand, more differentiated and highly proliferative hematopoietic progenitors favor OXPHOS[191]. In contrast, small intestinal stem cells favor OXPHOS independent of oxygen supply and against the general notion that adult stem cells favor glycolytic metabolism[187], while neighboring and terminally differentiated Paneth cells are characterized by glycolytic metabolism[192]. In addition, lactate (the waste product of glycolysis in Paneth cells) fuels OXPHOS in intestinal stem cells. Thus, metabolic heterogeneity can be an intrinsic feature

of cellular subtypes in normal tissues and is not only the consequence of environmental influences.

Aberrant proliferation of tumor cells requires the adaption of metabolism in order to sustain increased energy consumption. Otto Warburg first observed a counterintuitive effect in cancer cells, which mostly limit their energy metabolism to glycolysis even in the presence of oxygen[193], probably to generate glycolytic intermediates to fuel biosynthetic pathways[194]. However, recent studies revealed more complex metabolic heterogeneity similar to normal tissues. Beside strong microenvironmental influences on metabolic preferences in tumor cells[195], that could be also linked to WNT signaling in CRC[196], CSC features seem to be characterized by distinct metabolic states depending on the tumor entity[197]. For example, the acquisition of breast CSCs properties depends on a switch to glycolytic metabolism[198], while putative colorectal CSCs have increased mitochondrial function compared to tumor cells without CSC properties[199]. Furthermore, Sonveaux et al. revealed metabolic symbiosis of tumor cells based on lactate similar to those described for the normal intestinal stem cell niche[200].

Taken together, these results indicate that intrinsically regulated metabolic preferences and interdependencies during cancer cell differentiation might represent promising targets for cancer therapy. However, single cell metabolic states and underlying gene expression networks have been poorly defined. Although new methods for single cell metabolomics are emerging[201], scRNA-seq might help to reveal metabolic cancer cell heterogeneity based on the wiring of metabolic and transcriptomic networks[202]. In addition, imaging based *in situ* analysis will support the dissection of microenvironmentally influenced or differentiation related metabolic heterogeneity in primary samples or *in vitro*.

1.3.2 Analysis of tumor cell heterogeneity *in vitro*

Although *in vitro* cell culture systems are still limited in reflecting important stromal and microenvironmental characteristics of intratumor heterogeneity, they have proven highly valuable to understand important features of functional tumor cell heterogeneity. Advances in single cell analysis by imaging and NGS technology have provided further insight and will likely continue to be the basis for further studies. For example, bulk RNA-seq combined with multiplexed RNA-FISH revealed rare, transient transcriptional states that confer drug resistance in melanoma cells cultured in 2D[203]. Very recently, scRNA-seq combined with computational correction of dropouts has been used to study Transforming Growth Factor Beta (TGF-β) induced EMT in 2D cultured transformed mammary epithelial cells, revealing asynchronous induction and networks of transcriptional regulators that govern EMT[121].

Beyond these basic principles underlying EMT and drug resistance, patient-derived cultures require a more physiological environment to more closely reflect and maintain characteristics of the primary tumor, which is now achieved by 3D cell culture systems[17,32] (see section 1.1.3). Initially, single cell analysis of tumor cell heterogeneity in 3D cell culture systems was performed by isolating subpopulations by FACS based on a few defined markers, and their subsequent xenotransplantation[30,31]. Later on, genetic marking of subclones enabled *in vivo* tracing of single cell behavior. For example, lentiviral delivery of molecular barcodes into cells derived from CRC spheroids in combination with xenotransplantation and next generation sequencing (NGS) was used to trace the proliferative potential of single clones[204], which validated the presence of functionally distinct subpopulations *in vivo*. More recently, genetic engineering of organoids enabled lineage tracing and ablation of $LGR5^+$ cells upon xenotransplantation and revealed the functional role of $LGR5^+$ tumor cells in CRC[205] (see section 1.3.1.2). NGS has been primarily used for bulk expression profiling in 3D cell culture systems[34], but has been recently extended for single cell analysis. Roerink et al. used single cell isolation and clonal organoid expansion in combination with molecular profiling to investigate CRC evolution in single patients, thereby revealing extensive diversification at the DNA, methylome and transcriptome level that are maintained in 3D culture[206].

Although several single cell transcriptome studies could dissect subtype compositions in healthy small intestinal organoids[22,88,207], state-of-the-art single cell analysis in 3D tumor cell culture systems has not been performed in further detail. A deeper understanding of complex cellular phenotypes and underlying gene expression will be required to understand tumor cell heterogeneity as basis for functional studies in the future, especially because advanced *in vitro* systems are capable of closely reflecting *in vivo* characteristics.

1.4 Aim of study

Cellular behavior is the direct consequence of its underlying gene expression network, which itself is influenced by the genetic background, developmental programs and the cellular environment. A key goal in biology is to understand these gene(network) – function relationships and recent technical advances have now opened the door to characterize genomes, gene expression and epigenetic states in its native context: the single cell. Although transcriptomes can now be reliably measured at the single cell level in an unbiased way, these technologies require the dissociation of tissues into single cell suspensions which consequently results in the loss of contextual and phenotypic information. On the other hand, imaging-based methods for *in-situ* single cell analysis are generally restricted by the limited number of pre-selected markers that can be analyzed simultaneously. Moreover, hybrids of

imaging and sequencing are only beginning to emerge. Thus, direct or indirect combinations of both NGS and imaging technologies are required to fully understand contextual cellular phenotypes and underlying gene expression. Intratumor heterogeneity represents a key phenomenon in translational research whose mechanisms are poorly understood, and recent studies have impressively demonstrated the high cellular complexity in various tumor entities by single cell analysis and sequencing. Still, very little is known about functional tumor cell heterogeneity including subtype-specific behavior, spatial organization and cellular interactions. As the analysis of contextual single cell phenotypes in primary tumors is limited to histology-based methods, functional assays based on 3D *in vitro* cell culture systems that closely mirror tumor subtype compositions and neoplastic tissue architecture are required.

In order to understand intratumor heterogeneity beyond subtype composition, this study aims to provide a framework for 3D *in vitro* analysis of tumor cell heterogeneity by directly and indirectly combining NGS and quantitative microscopy. Therefore, we use, adapt and extend existing state-of-the-art sequencing and imaging technology and apply these to established and patient-derived 3D cell culture systems, which provide the optimal trade-off between flexibility and complexity for single cell analysis.

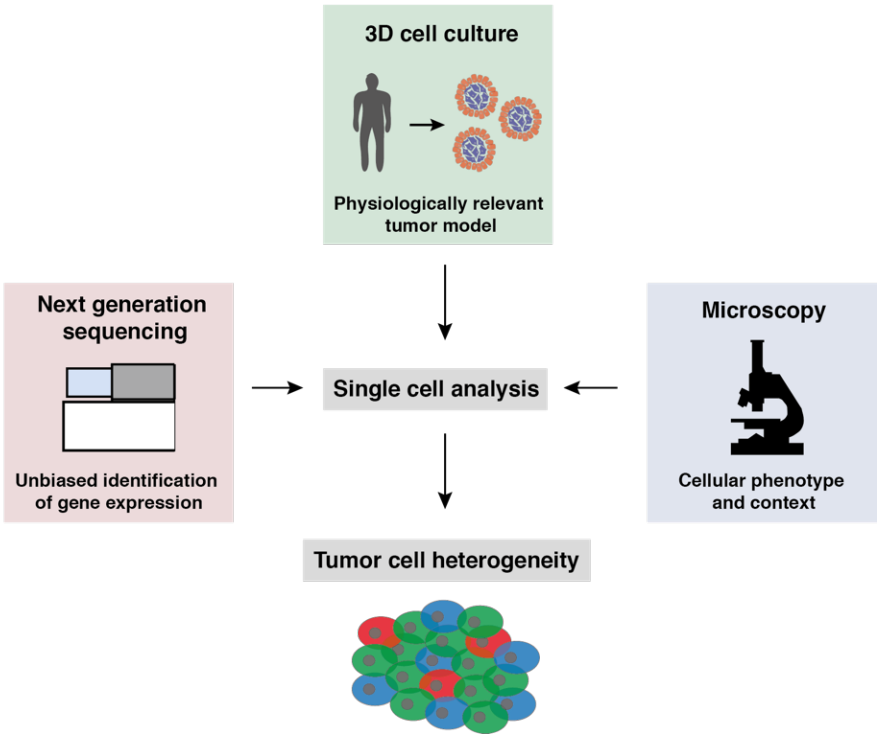


Figure 1.6 | Aim of this study as graphical overview. 3D cell culture systems provide the physiological context for *in vitro* culture of patient-derived material, NGS enables unbiased identification of gene expression and microscopy gives information about cellular phenotypes and context. This study aims to use, adapt and improve state-of-the-art technology to directly and indirectly combine NGS and quantitative microscopy for single cell analysis to understand tumor cell heterogeneity in 3D cell culture systems.

2 Results

This results chapter is subdivided into two major sections. First, I describe the development of the 'pheno-seq' method that directly combines high-throughput imaging and gene expression profiling of clonal spheroids and its application on established and patient-derived 3D culture models of breast and colorectal cancer (section 2.1). Second, I will present the results of dissecting the cellular composition of 12 spheroid culture lines derived from patients with CRC by scRNA-seq. In addition, I will present the validation and extension of sequencing results by multiplexed RNA-FISH analysis *in situ* (section 2.2). For data generated or analyzed in collaboration with others, co-workers are indicated by name in figure legends. More detailed information about contributions can be found on page iii. Figures and text in section 2.1 are widely adapted from the associated publication for which I have written the original text[208]. Although multicellular structures derived from MCF10CA breast cancer cells could, by definition[209], also be described as organoids, I use the term 'spheroid' throughout the whole work for both CRC and MCF10CA.

2.1 Pheno-seq – linking morphological and functional features to gene expression in 3D cell culture systems

2.1.1 Using single cell *in-vitro* 3D cell culture to analyze patho-phenotypes of tumor cells

Although single cell gene expression profiling provides biologically rich information, identified genes that can be assigned to subtypes do not necessarily inform about complex cellular phenotypes. However, visual characteristics of cells or cell clusters can be highly informative, especially for classification of tumor subtypes and disease states[210]. This also holds true for patient-derived 3D cell culture systems, but most studies have so far focused on inter-patient differences[35,211] rather than heterogeneous phenotypes and behavior of cells isolated from a single patient.

Single-cell 3D-culture in combination with microscopy and molecular analyses appears as a key strategy for the analysis of functional tumor cell heterogeneity *in-vitro* as it enables analysis of clonal behavior in defined spatial and temporal conditions[206,212]. Ideally, the visual phenotype of the cell or the emerging multicellular complex (spheroids, organoids, etc.) reflects the characteristics of the primary neoplastic tissue and consequently informs about the functional outcome of heterogeneous cancer cell states. Informative subpopulation-specific oncogenic phenotypes can reflect differences in long-term proliferative capacity[204],

or morphologically complex phenotypes such as deregulation of epithelial growth and invasiveness, a well-established prerequisite for metastasis[180].

In breast cancer entities that originate from the mammary gland, normal epithelial cells undergo a stepwise transformation from local hyperplasia to premalignant carcinoma *in-situ* and invasive carcinoma[213] (for further description of cancer cell invasion see section 1.3.1.4). Importantly, the switch from epithelial to invasive phenotypes requires transcriptional programs that resemble those occurring during embryogenesis and wound healing, commonly described as epithelial-to-mesenchymal transition (EMT)[180]. A common 3D model of tumorigenic progression in breast cancer is the basal-like MCF10 progression line[214,215]. Originating from the non-neoplastic immortalized parent cell line MCF10A[216], more transformed derivatives have been generated by transfection with the c-Ha-ras oncogene and by xenograft passaging[217–219]. Of those, the MCF10CA cell line represents a fully malignant derivative with invasive and metastatic properties in xenografts[218].

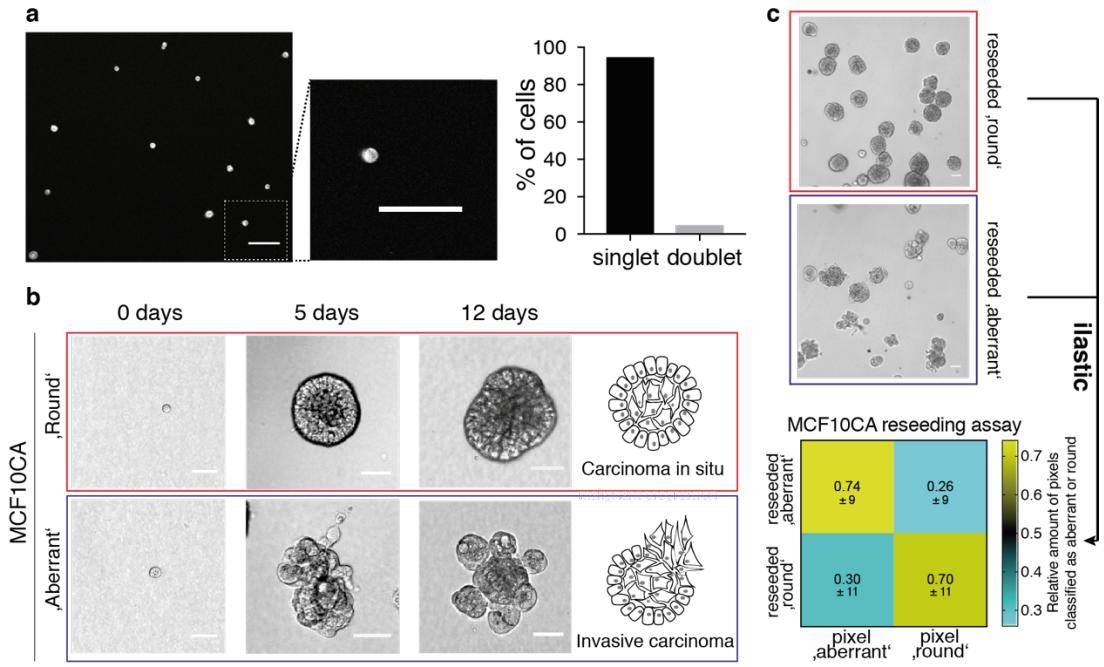


Figure 2.1 I Breast cancer 3D model MCF10CA. (a) Single-cell seeding efficiency of MCF10CA cells assessed by image analysis. Example image of CellTracker Red stained and seeded cells and magnified image that corresponds to dashed box in left image (scale bar: 100 μm). Right bar plot: Cell singlets and doublets imaged and quantified after seeding (289 objects in total). **(b)** Brightfield microscopy images of clonal MCF10CA spheroids in Matrigel after 0, 5 and 12 days of culture. Spheroid phenotypes reflect histological characteristics of key steps during malignant progression of breast cancer (Brightfield, scale bar 50 μm). Red box: 'round' phenotype; Blue box: 'aberrant' phenotype. **(c)** Spheroids derived from cells independently isolated from 'round' and 'aberrant' spheroid phenotypes and quantification after regrowth by 'ilastic' machine learning based pixel classification. Upper: Example images of reseeded MCF10CA 'round' and 'aberrant' spheroids 5 days after reseeding (scale bar: 50 μm). Lower: Spheroid classification confusion matrix. Heatmap reflecting classified pixels by ilastic as aberrant or round after reseeding (four replicates, indicated are relative pixel numbers and standard error of the mean below). Image analysis in (b) and (c) has been done together with Friedrich Preußer.

As invasive properties are most likely not inherent to all cells from a tumorigenic epithelial cell line, we expected heterogeneous cell states in the MCF10CA cell line with different phenotypic characteristics. In order to follow growth of individual cells, we first established a simple single cell seeding strategy in reconstituted basement membrane (Matrigel) and assessed seeding efficiency by image analysis (Figure 2.1a, Figure 2.2a). After 5 days in 3D culture, single-cell-derived MCF10CA spheroids show a remarkable morphological heterogeneity, with cellular phenotypes reflecting characteristics of both carcinoma *in-situ* ('round' phenotype) and invasive carcinoma ('aberrant' phenotype) (Figure 2.1b). Next, we developed a workflow based on enzymatic digestion to isolate single spheroids without perturbing their phenotypic identity (Figure 2.2b) in order to analyze cells derived from both phenotypes independently. To functionally assess the observed visual heterogeneity, we reseeded and cultured cells from both phenotype classes independently and quantified reoccurrence of spheroid phenotypes by supervised machine learning[220]. Results inferred by this strategy revealed a high cell state stability and validated efficient isolation of spheroid phenotypes (Figure 2.1c).

As local invasion of cancer cells and the formation of distant metastases are critical events during progression to higher pathological grades of malignancy, we reasoned that the MCF10CA cell line represents a valid proof-of-concept model to further analyze heterogeneous and pathologically relevant spheroid phenotypes.

2.1.2 Pheno-seq as new approach to relate clonal spheroid phenotypes to gene expression

As next step, we aimed to understand the link between heterogeneous spheroid phenotypes and associated changes in gene expression. Based on a commercial continuous-flow microfluidic platform (Fluidigm C1)[79], we first generated and deeply sequenced full-length scRNA-seq libraries of both 'aberrant' and 'round' phenotypes independently (166 cells in total, Figure 2.2c). Notably, this strategy does not enable a direct phenotypic correlation as multiple spheroids (>30) needed to be pooled and dissociated to ensure a sufficient number of input cells. For transcriptomic analysis of cells from both spheroid phenotypes combined, we tested annotated and *de-novo* identified gene sets for coordinated expression variability across cells[116]. After correcting for cell cycle variability, tSNE[112] 2D embedding revealed two distinct clusters and a tight association of cells to their original spheroid phenotype class (Figure 2.2d).

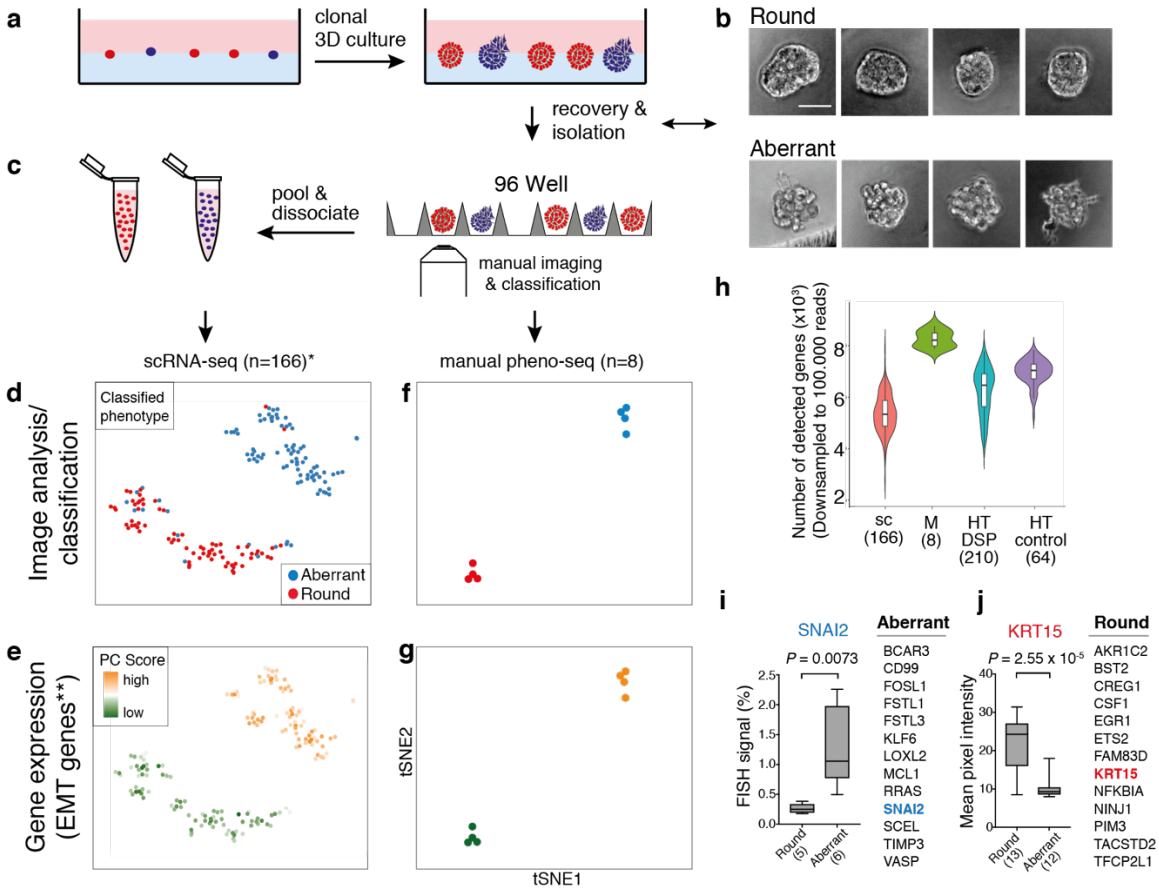


Figure 2.2 I Pheno-seq enables direct image correlation and complements the identification of morphology-specific gene expression. (a) Workflow overview for the isolation of clonal spheroids for the identification of morphology-specific gene expression. **(b)** Brightfield images of clonal MCF10CA spheroids (phenotype classes 'round' and 'aberrant') after isolation from Matrigel (scale bar 50 μm). **(c)** Indirect phenotype – transcriptome correlation by scRNA-seq using cells isolated from multiple (>30) spheroids with annotated morphology phenotype. **(d)** tSNE visualization[116] of 166 scRNA-seq (*cell-cycle corrected) full-length expression profiles of cells derived from manually isolated round and aberrant spheroids. Coloring based on manual phenotype classification. **(e)** Same tSNE visualization as shown in (d) with coloring based on PC scores for **HALLMARK_EMT gene set derived from the Molecular Signature Database[221] (MSigDB). **(f and g)** tSNE visualization of 8 full-length manual pheno-seq expression profiles based on manually isolated single spheroids. Same coloring as presented in (d) and (e). **(h)** Number of detected genes in downsampled scRNA-seq and pheno-seq libraries (sc: scRNA-seq; M: manual pheno-seq; HT-DSP: high-throughput pheno-seq combined with dithio-bis(succinimidyl) propionate fixation; HT-control: HT-pheno-seq bottom control). Numbers of samples indicated on x-axis under respective strategy. **(i and j)** Selected genes identified by manual pheno-seq and not by scRNA-seq (Differential expression analysis[109]: Fold change > 1.3; adjusted P-value < 0.1) and imaging based validation of phenotype-specific expression for SNAI2 (aberrant) and KRT15 (round). RNA-FISH for SNAI2: Plotted values reflect the percentage of pixels that exceed the background threshold per spheroid. KRT15 immunofluorescence: Plotted values reflect the mean pixel intensity per spheroid. Box plot center-line: median; box limits: first and third quartile; whiskers: min/max values. Numbers of samples indicated on x-axis under respective phenotype class. Indicated are *P*-values from unpaired two-tailed Students t-test. scRNA-seq libraries shown in (d) and (e) have been generated together with Jan-Philipp Mallm. RNA-seq analysis in (d-h) has been performed together with Jeongbin Park, Simon Steiger and Zuguang Gu. Image analysis in (i) has been done together with Friedrich Preußer.

Furthermore, differential expression analysis[109] identified biologically relevant expression patterns: Cells derived from aberrant spheroids show enhanced expression of EMT related genes (Figure 2.2e, Figure 2.3a, Supplementary Figure 1a), including vimentin (VIM), Beta-

Actin (ACTB) and fibroblast activating protein (FAP), whereas cells isolated from round phenotypes showed higher expression of genes involved in adherence and formation of tissue structures including desmoglein 3 (DSG3) and keratin 16 (KRT16) (Figure 2.3a). Next, we validated RNA-seq results by whole mount immunofluorescence (IF) of individual marker genes for aberrant phenotypes, in particular the EMT marker VIM and the cytoskeleton component ACTB (Supplementary Figure 2a and b).

Current scRNA-seq methods can be strongly affected by low RNA input[72] and dissociation bias[102]. To avoid these technical influences, we next tested expression profiling of manually isolated single spheroids (manual pheno-seq) as complementary approach to identify transcriptional differences between clonal spheroid phenotypes. Despite the loss of single-cell resolution, we reasoned that pheno-seq should improve accuracy by enabling a direct correlation of image phenotype to transcriptome, and at the same time provide more RNA material for cDNA library preparation. First, we started with low spheroid sample numbers in a tube-based setup to evaluate the ability to detect relevant heterogeneous gene expression that is missed by scRNA-seq. Manual pheno-seq expression profiling of only eight spheroids yielded a similar phenotype-specific clustering defined by high and low expression of EMT-related genes (Figure 2.2f and g, Figure 2.3a, Supplementary Figure 1b). Although the sample number was approximately 20 times lower (166 single-cells vs. 8 single spheroids), the gene detection rate per sample was significantly higher compared to scRNA-seq (Figure 2.2h, Supplementary Table 1), and differential expression analysis revealed over 50 phenotype-specific genes for each of the two phenotype classes that could not be detected by scRNA-seq (Figure 2.2 i and j, Figure 2.3b). Importantly, these genes include the transcriptional EMT master regulator SNAI2[222] (aberrant) and keratin 15 (KRT15, round) a basal-myoepithelial marker in the mammary gland[223] (Figure 2.3c, Supplementary Figure 1b). However, we detected more differentially expresses genes by scRNA-seq, which is most likely due to the much higher sample number.

We validated spheroid phenotype-specific expression of SNAI2 and KRT15 by RNA-FISH and immunofluorescence (IF), respectively (Figure 2.2 i and j, Supplementary Figure 2c and d). We reasoned that SNAI2 was not identified by scRNA-seq due to its low expression (Figure 2.3c), a frequent phenomenon for transcriptional regulators in EMT[121]. Although KRT15 is one of the top pheno-seq markers for round spheroids, the presence of residual $KRT15^{+}$ cells in aberrant spheroids (Supplementary Figure 2c) seemed to mask the identification of KRT15 as round-specific when single-cell profiles were analyzed. Remarkably, differential expression of KRT15 and SNAI2 could not be robustly restored from single cell data by generating pseudo pheno-seq profiles from averaged scRNA-seq

expression (Figure 2.3d, Supplementary Figure 1c), indicating for the additional influence of dissociation bias on KRT15 mRNA abundance. In sum, pheno-seq provides the direct correlation of clonal spheroid phenotypes and transcriptomes and complements scRNA-seq in identifying heterogeneous gene expression already with low sample numbers.

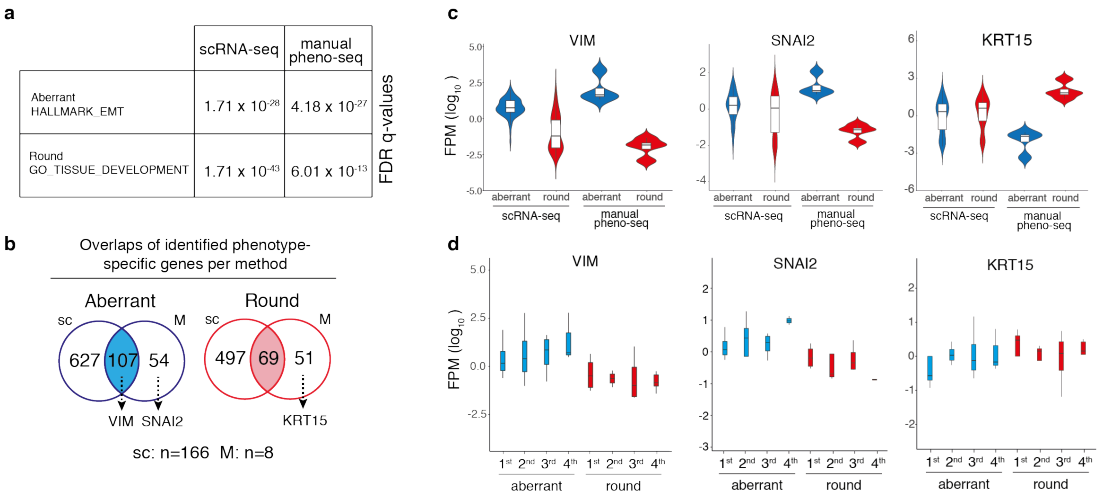


Figure 2.3 I pheno-seq identifies highly relevant gene expression that is missed by scRNA-seq. (a) Gene set enrichment analysis of differentially expressed genes identified by scRNA-seq and manual pheno-seq. Shown are FDR q-values for enrichments of HALLMARK_EMT and GO_TISSUE_DEVELOPMENT gene sets (derived from the MSigDB). **(b)** Venn-Diagrams showing overlaps of identified phenotype-specific genes between scRNA-seq and manual pheno-seq identified by differential expression analysis (fold change>1.3; adjusted p-value<0.1). **(c)** Violin plots presenting expression of individual genes (VIM, SNAI2, KRT15) for identified phenotype-specific clusters for scRNA-seq and manual pheno-seq. Expression magnitude is plotted as Fragments per Million (FPM, log_{10}). Violin-plot center-line: median; box limits: first and third quartile; whiskers: ±1.5 IQR. **(d)** Boxplots reflecting expression of individual genes (VIM, SNAI2, KRT15) per phenotype-specific clusters for pseudo pheno-seq profiles. Expression magnitude is plotted as Fragments per Million (FPM, log_{10}) of four independent randomizations. Boxplot center-line: median; box limits: first and third quartile; whiskers: ±1.5 IQR. RNA-seq analysis in (b-d) has been done together with Jeongbin Park.

2.1.3 Development of high-throughput pheno-seq in barcoded nanowells

A key limitation of both scRNA-seq and manual pheno-seq presented in the previous section is the non-quantitative and biased selection of spheroid phenotypes based on visual inspection by eye. In addition, profiling a higher number of spheroids per pheno-seq experiment is necessary to comprehensively understand the link between visual phenotypes and gene expression in 3D culture models. Therefore, we developed high-throughput (HT) pheno-seq by adapting and improving the nanowell-based iCELL8 scRNA-seq system[224], a technology for integrated imaging and gene expression profiling of single cells or nuclei, for transcriptomic profiling of spheroid samples of up to 100 μm in size. Major modifications for MCF10CA spheroids included: (i) cellular fixation[225] compatible with RNA isolation and

reverse transcription in order to compensate for prolonged imaging time, (ii) altered chip setup and higher-resolution confocal microscopy for detection of complex spheroid phenotypes, (iii) an automated image-processing pipeline (iv) and the 'PhenoSelect' software for interactive analysis and selection of spheroids for sequencing (Figure 2.4, Supplementary Figure 3, Supplementary Figure 4). These significant technical changes had only minor influences on the gene detection rate, which fell in between scRNA-seq and manual pheno-seq (Figure 2.2h, Supplementary Table 1).

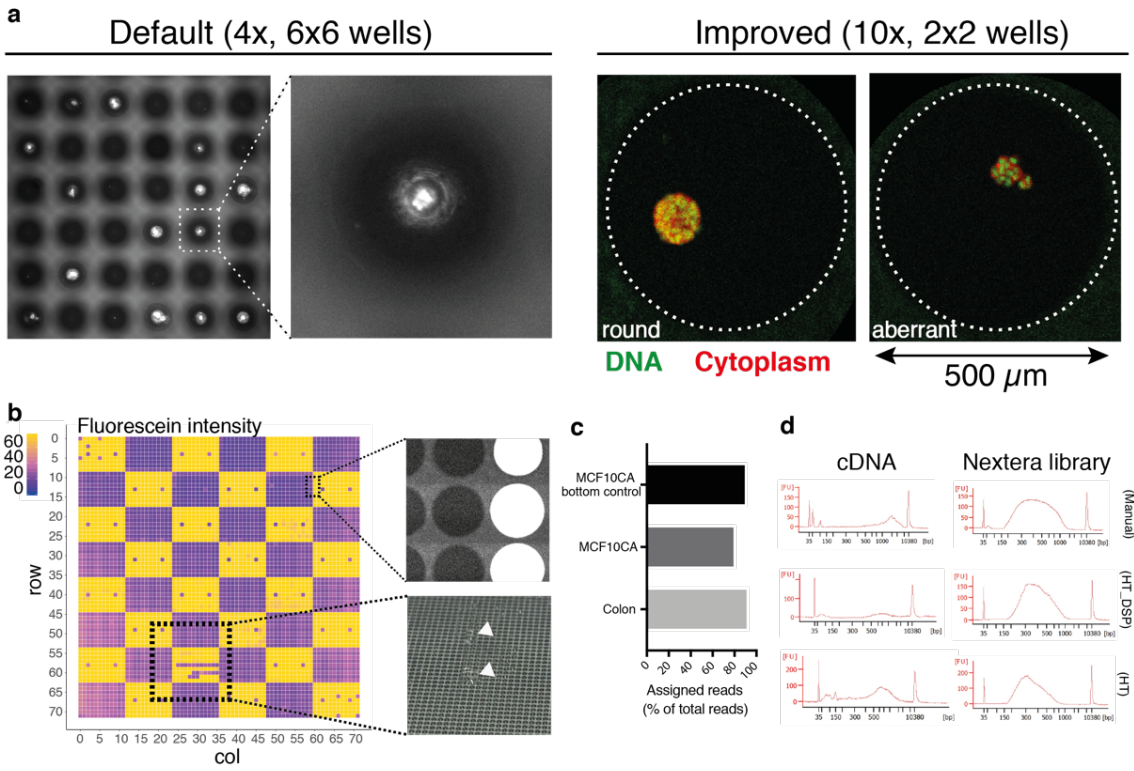


Figure 2.4 I Technical adaptions and controls for high-throughput pheno-seq. (a) Comparison of images acquired by the default iCELL8 microscope (4x objective, 6x6 wells per image) with spheroid nuclei stained with Hoechst dye, and higher resolution microscopy with confocal laser scanning microscope (10x objective, 2x2 wells per image) with spheroids stained with Hoechst dye and CellTracker Red CMTPX. **(b)** Patterned Fluorescein dispensing for leakage analysis. Average fluorescence intensity plotted onto 72x72 well grid corresponding to nanowell chip architecture (left). All average intensity values exceeding 77 were set to maximum in the color code scheme for better visualization. Top right: Image example showing border between wells that have been filled with PBS or PBS with Fluorescein. Lower right: Macroscopic image of nanowell surface with droplets, showing dispensing errors that are reflected by the absence of fluorescence signal at the associated position. **(c)** High percentage of reads that only map to selected well barcodes excludes significant leakage of barcoded Poly-T primers upon centrifugation of spheroids to the foil. **(d)** cDNA and Nextera XT sequencing library Bioanalyzer traces show compatibility of HT-pheno-seq with iCELL8 system. HT pheno-seq microscopy in (a) and image analysis in (b) has been done together with Friedrich Preußer. RNA-seq analysis in (c) has been done together with Jeongbin Park.

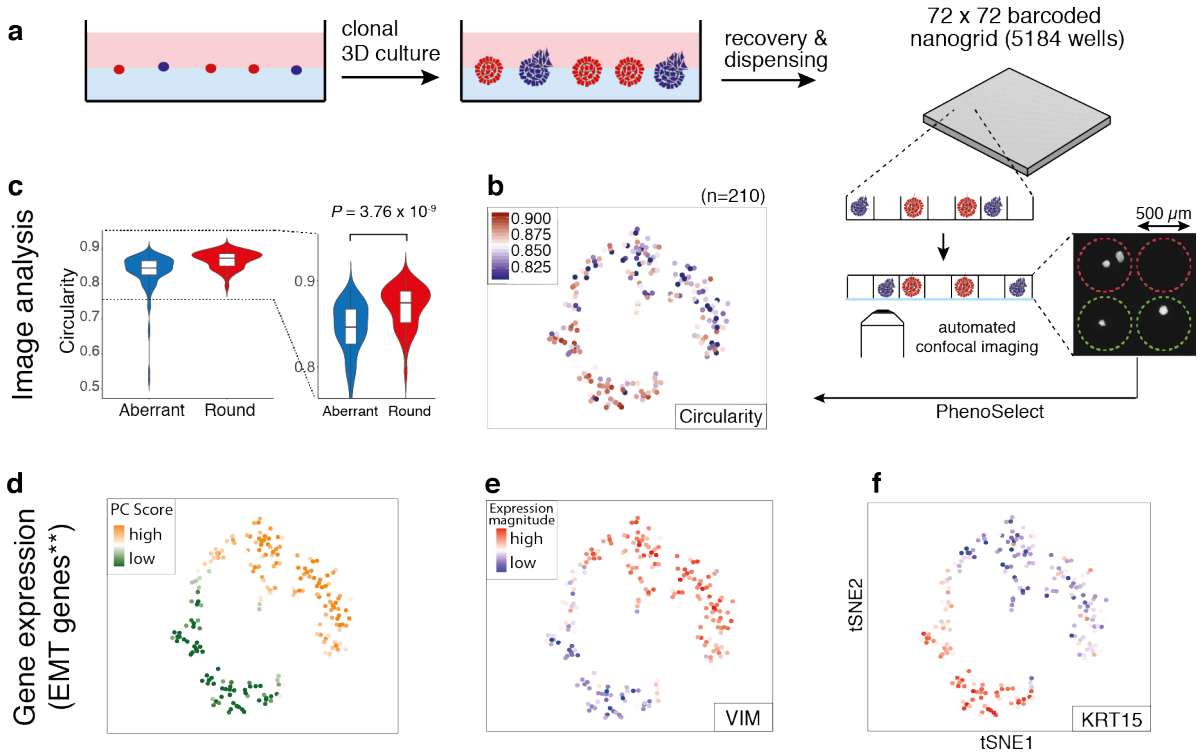


Figure 2.5 I High-throughput pheno-seq of MCF10CA spheroids. (a) High-throughput (HT) pheno-seq workflow for MCF10CA spheroids based on automated dispensing and confocal microscopy of recovered spheroids in barcoded nanowells. **(b)** tSNE visualization of 210 HT-pheno-seq 3´-end profiles with individual spheroid data points colored by image feature 'circularity'. For better visualization, all circularity values below 0.8 were set to minimum in the color code scheme. **(c)** Circularity plotted per cluster (k-means clustering, k=2) as shown in (b). Violin-plot center-line: median; box limits: first and third quartile; whiskers: ±1.5 IQR. Indicated P-value derived from unpaired two-tailed Students t-test. **(d)** Same tSNE visualization as shown in (b) but coloring based on PC scores for **HALLMARK_EMT gene set. **(e and f)** Same tSNE visualization as shown in (b and d) but coloring based on expression magnitude for genes VIM (e) and KRT15 (f). RNA-seq and image analysis in (b-f) has been done together with Jeongbin Park. The automated microscopy workflow, the image pre-processing pipeline and PhenoSelect have been jointly developed with Friedrich Preußer.

MCF10CA HT-pheno-seq yielded very similar results as described for manual pheno-seq in the previous section, with two distinct clusters driven by expression of genes involved in EMT (VIM$^+$) as well as tissue formation (KRT15$^+$) but at much higher throughput (n = 210) (Figure 2.5a, d, e, f). Both pheno-seq approaches show good concordance in identifying differentially expressed genes between spheroid phenotypes (Supplementary Figure 5c), despite unbiased capture of spheroids by HT-pheno-seq as well as differences in sample number and library structure (3´-end vs. full-length, Supplementary Table 1).

HT-pheno-seq allows profiling of mRNA abundance and image features from the same spheroid, which enabled straightforward correlation of genetic programs and complex visual phenotypes based on the fluorescence signal emitted from a cytoplasmic dye (CellTracker Red). Biologically relevant phenotypes included the morphology-related feature 'circularity' which informs about (de)regulation of lobular development (Figure 2.5b and c), and spheroid

size, indicating for a higher proliferative activity of epithelial cells (Supplementary Figure 5a). In addition, pheno-seq linked negatively skewed pixel intensity distributions to round phenotypes (Supplementary Figure 5b), indicative of an increased cell density in round spheroids that leads to an increased fraction of high pixel intensity values derived from the cytoplasmic signal. Hence, HT-pheno-seq represents a new method that, unlike scRNA-seq, directly and quantitatively links heterogeneous visual phenotypes to underlying gene expression in a single experiment.

2.1.4 HT-pheno-seq of a patient-derived 3D model of colorectal cancer

Next, we set out to analyze the functional correlation between visual phenotypes and gene expression in a physiologically relevant and patient-derived 3D model isolated from a liver metastasis of a CRC patient (Figure 2.6a). Similar to the phenotypic heterogeneity in the MCF10CA model described in the previous section, functionally distinct subpopulations in 3D cultures of CRC patients have been previously identified[204]. The reported heterogeneity in proliferative capacity of single cells seems to be independent of mutational subclone diversity[226], thereby supporting the existence of a differentiation-like hierarchy in CRC (see section 1.3.1.2). As reseeding of cells from different spheroid sizes classes (20-40 μm and >70 μm) revealed substantial differences in spheroid forming capacity (Figure 2.6b), we reasoned that specific stem- and differentiation-related gene expression signatures should underlie these heterogenous proliferative phenotypes.

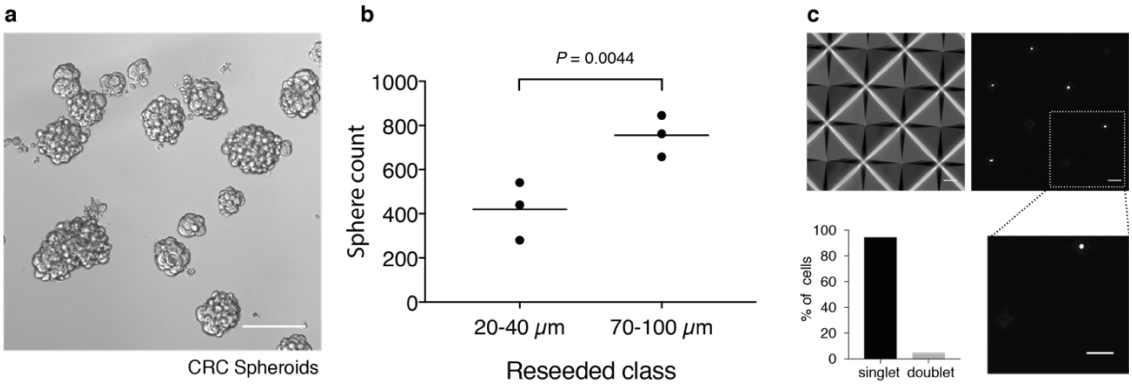


Figure 2.6 I 3D spheroid model of colorectal cancer. (a) Brightfield microscopy image of 10 day cultured clonal CRC spheroids derived from a liver metastasis (scale bar 100 μm). **(b)** Reseeding assay with cells isolated from distinct spheroid size classes (20-40 μm and >70-100 μm). Plotted are spheroid counts 10 days after reseeding (three replicates, center-line: mean; indicated P-value of paired two-tailed Students t-test). **(c)** Single-cell seeding efficiency in inverse pyramidal shaped microwells (upper left) assessed by image analysis. Upper right: Example image of CellTracker Red stained cells seeded in microwells (scale bar: 100 μm). Lower right: Magnified image that corresponds to dashed box in upper right image. Lower left: Quantified cell singlets and multiplets after seeding (three wells, four images per well, 70 objects in total). The CRC spheroid culture has been provided by Hanno Glimm and Claudia Ball. Image analysis in (b) and (c) has been jointly performed with Friedrich Preußer.

2.1.4.1 Analysis of relative transcript abundances between CRC spheroids

To investigate this hypothesis, we performed HT-pheno-seq based on clonal CRC spheroids cultured in an inverse pyramidal-shaped microwell setup (Figure 2.6c, Figure 2.7a). 2D tSNE visualization of relative gene expression differences between 95 HT-pheno-seq profiles revealed two transcriptionally distinct clusters (Figure 2.7b) and image analysis of the associated spheroids showed a strong difference in spheroid size composition between both clusters (Figure 2.7b and c).

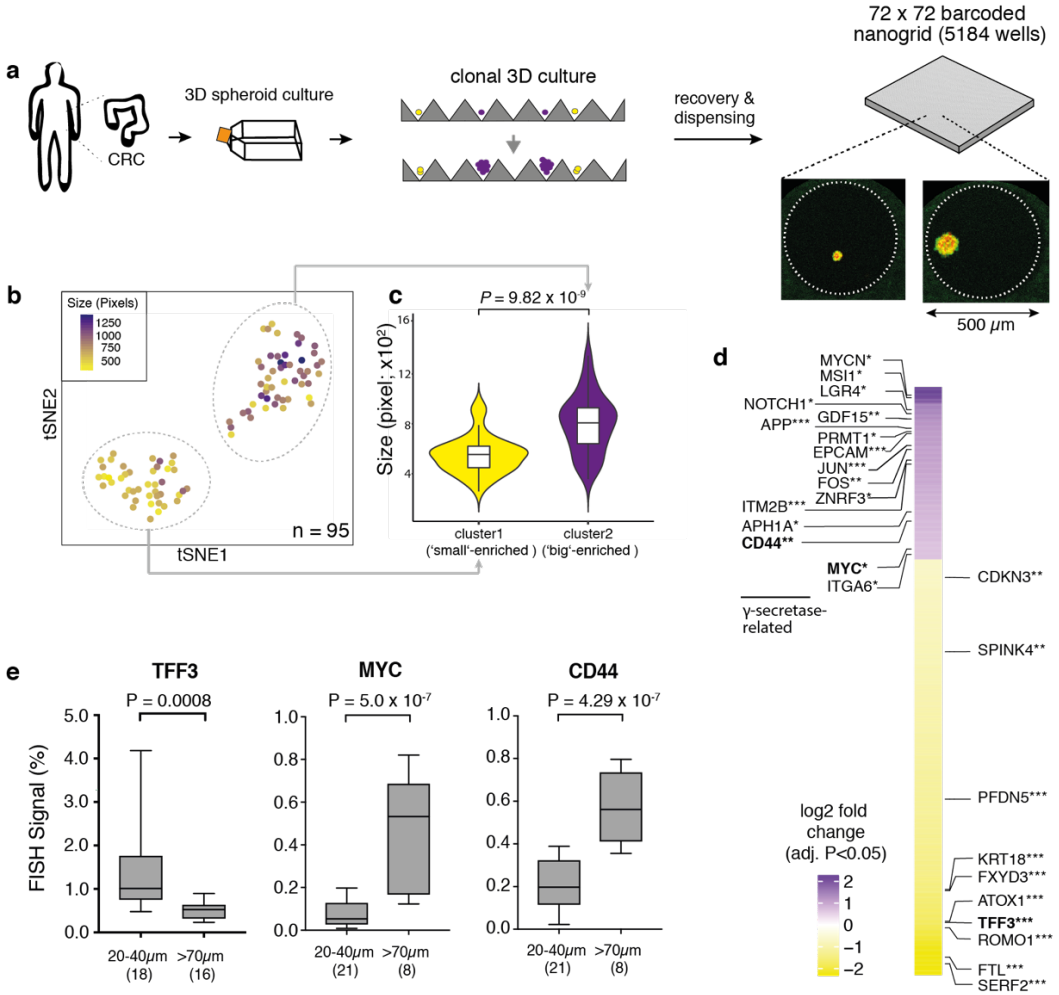


Figure 2.7 I HT-pheno-seq of a 3D model of colorectal cancer links heterogeneous proliferative phenotypes to expression signatures enriched for cell type-specific markers. (a) Clonal 3D culture in inverse pyramidal shaped microwells and recovery for HT-pheno-seq of CRC spheroids. Yellow and purple cells reflect heterogeneous subpopulations with functional differences in long-term proliferative capacity. **(b)** 2D tSNE visualization of 95 HT-pheno-seq gene expression profiles. Coloring by spheroid size (in pixels). **(c)** Violin plot describing spheroid sizes per cluster as shown in (b). Violin-plot center-line: median; box limits: first and third quartile; whiskers: ±1.5 IQR). Indicated *P*-value calculated from unpaired two-tailed Students t-test. **(d)** Heatmap describing results of differential expression analysis of identified clusters in (b); purple: big-enriched, yellow: small-enriched. Selected genes are indicated beside the heatmap; Fold change > 1.5; adjusted P-value < 0.05; *P < 0.05, **P < 0.01, ***P < 0.001; cluster1 ('small-enriched'): 313 differentially expressed genes; cluster2 ('big' enriched): 130 differentially expressed genes. **(e)** Validation of HT-pheno-seq results by quantitative RNA-FISH for size-dependent differentiation marker TFF3 ('small') and stem cell markers CD44/MYC ('big'). Plotted values represent pixel fraction that exceeds the background threshold per spheroid (Box plot center-line: median; box limits: first and third quartile; whiskers: min/max values; *P*-values from unpaired Students t-test. Numbers of samples n indicated on x-axis under respective class). RNA-seq analysis in (b-d) has been performed together with Jeongbin Park. Image analysis in (e) has been done together with Friedrich Preußer

Differential expression analysis between detected clusters showed that the first cluster (‘small’ phenotype) is enriched for genes involved in ribosomal activity (GO_RIBOSOME, FDR q-value 2.41×10^{-45}) as well as for intestinal secretory lineage markers, including Trefoil Factor 3 (TFF3), KRT18 and SPINK4[88] (Figure 2.7d). In contrast, the second cluster (‘big’ phenotype) is characterized by the expression of genes previously described to be involved in (i) stem cell maintenance (including CD44, MYC, NOTCH1, APP, MSI1 and ITGA6)[88227], (ii) the formation of cell-cell junctions (including EPCAM, CLDN4, CDH1) and (iii) WNT signaling (ZNRF3, LGR4, JUN) (Figure 2.7d).

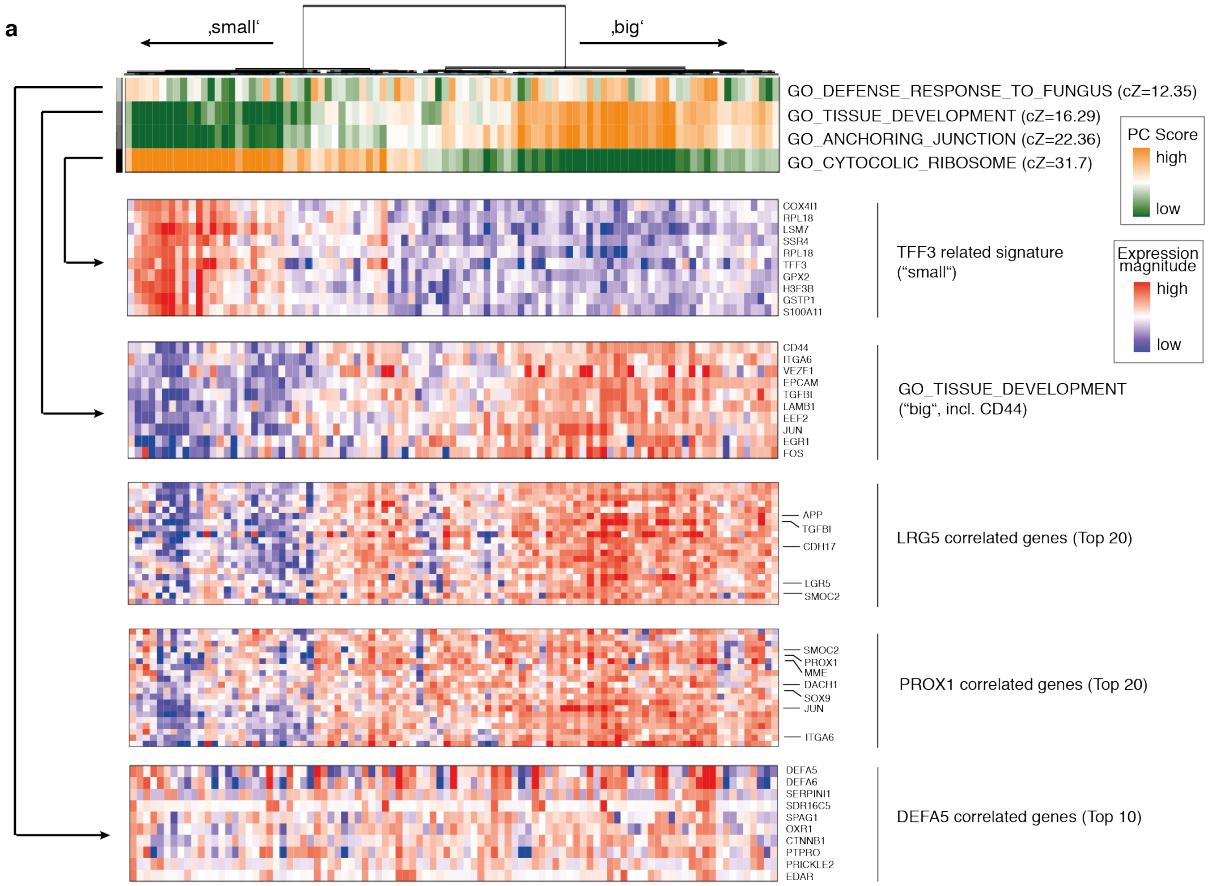


Figure 2.8 I Identified pheno-seq expression signatures for CRC spheroid model. PAGODA RNA-seq analysis of CRC spheroid HT-pheno-seq data. Dendrogram indicates overall clustering (left: ‘small’, right: ‘big’) and the rows below represent top four significant aspects of heterogeneity based on HALLMARK/GO gene sets derived from the MSigDB and on de-novo identified gene sets. High aspect scores (PC Scores) correspond to high expression of associated gene sets. Corresponding top gene sets are listed next to rows (including cZ scores as measure of gene set overdispersion). Expression patterns below reflect top 10 loading genes for selected gene sets that are associated with respective aspects. Expression patterns of genes exhibiting the highest correlation to the major intestinal stem cell marker LGR5 and putative cancer stem cell marker PROX1 (Pearson’s correlation, top 20 genes). Bottom: Expression pattern of top 10 genes most highly correlated with Paneth marker DEFA5 is independent of the major (size-associated) clustering. RNA-seq analysis has been performed together with Jeongbin Park.

Furthermore, this expression signature showed a very high overlap with the top correlated genes of the major intestinal stem cell marker LGR5, including CD44, APP and SMOC2 (Figure 2.8). We validated sphere size-dependent expression for selected markers by quantitative RNA-FISH (Figure 2.7e, Supplementary Figure 6).

In the cluster enriched for big spheres, we identified several genes related to the γ-secretase machinery (Figure 2.7d), a key component of the Notch signaling pathway and target of novel therapies that aim to disrupt cancer stem cell signaling[228]. Selective targeting of the γ-secretase with a small molecule inhibitor in concentration ranges that have been shown to force colonic stem cells into differentiation[229] showed a pronounced inhibitory effect on spheroid growth (Figure 2.9). This finding suggests similar signaling dependencies of the normal and transformed intestinal stem cell niche and shows the potential of pheno-seq to identify relevant signaling components required for long-term cellular proliferation.

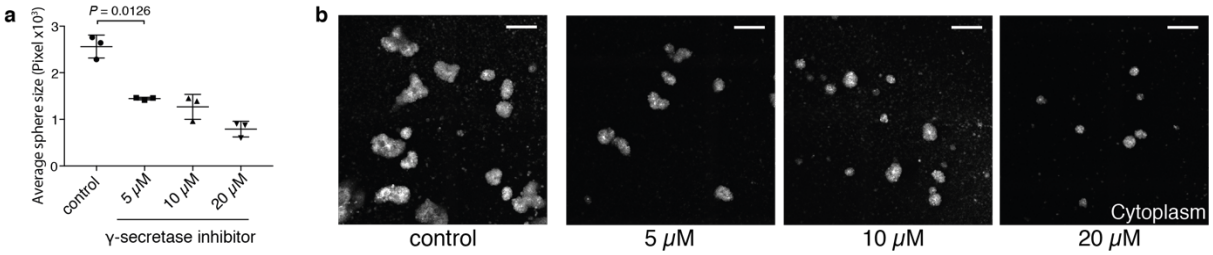


Figure 2.9 | Influence of γ-secretase inhibitor on spheroid growth. (a) Plotted values show average CRC spheroid sizes after 10 days in culture in the presence of different concentrations of the γ-secretase inhibitor PF-03084014 (Three replicates; dot plot center line: mean; whiskers: standard deviation; P-values from paired two-tailed Students t-test). **(b)** Example images of CellTracker Red stained spheroids in the presence of different γ-secretase inhibitor concentrations after 10 days in culture (scale bar 200 μm). Image analysis in has been done together with Marcel Waschow.

Moreover, we determined a transcriptional signature primarily driven by the expression of deep crypt secretory (Paneth) cell markers DEFA5 and DEFA6 that seems to be independent of the size-related clusters shown above (Figure 2.8). Paneth cells represent a post-mitotic secretory and anti-microbial subpopulation at the bottom of colonic crypts that serves as niche for LGR5⁺ stem cells[229]. In line with pheno-seq results, we validated high-expressing DEFA5⁺ cells as rare subpopulation with spheroid size-independent relative expression by RNA-FISH (Figure 2.10). Thus, pheno-seq is able to directly assign heterogeneous proliferative phenotypes to expression signatures enriched for specific intestinal cell-type markers, results that cannot be directly obtained from scRNA-seq data and also not without explicit single cell culture.

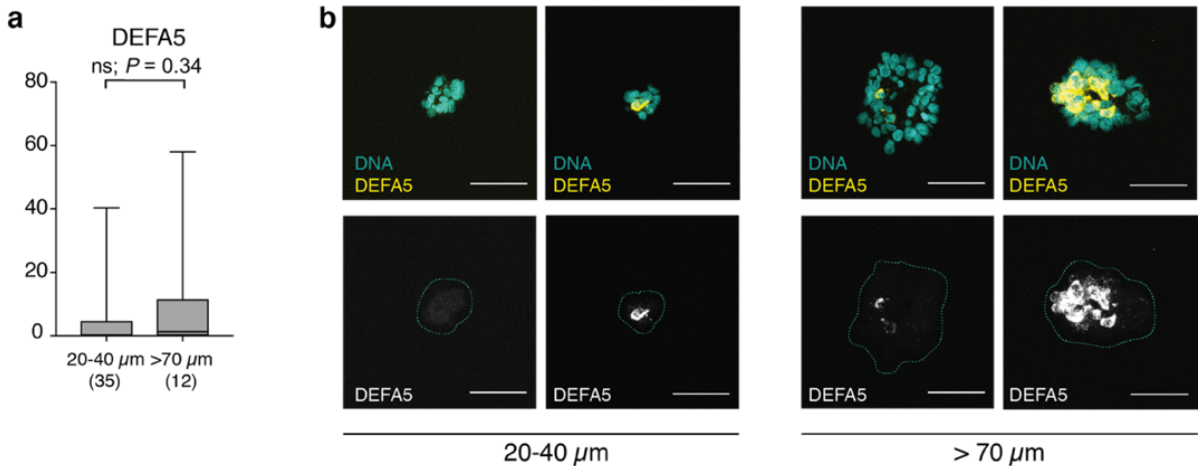


Figure 2.10 | DEFA5+ cells show heterogeneous growth phenotype. (a) Validation of pheno-seq by quantitative RNA-FISH for size-independent Paneth-cell marker DEFA5. Plotted values reflect the pixel fraction that exceeds the background threshold per spheroid (Box plot center-line: median; box limits: first and third quartile; whiskers: min/max values; *P*-values from unpaired Students t-test, ns: non-significant. Numbers of samples n indicated on x-axis under respective spheroid size class). **(b)** Example images shown as Z-projections for RNA-FISH staining for DEFA5 of big (>70 μm) and small (20-40 μm) spheroids with (top) and without (lower) Hoechst counterstain visualization (Hoechst: cyan; RNA: yellow). Dashed line in images without Hoechst visualization reflect spheroid border (scale bar 50 μm). Image analysis in has been done together with Friedrich Preußer.

2.1.4.2 Single cell deconvolution by image analysis and maximum likelihood inference

Pheno-seq enables the direct association of spheroid phenotypes and transcriptomes at a depth that cannot be reached by current scRNA-seq methods alone. However, these advances come at the cost of lower cellular resolution. Thus, gene expression signatures derived from CRC spheroids inform about general phenotype-specific expression and trends in subtype composition but might derive from multiple cellular subtypes present within the same spheroids. While these results are highly valuable for understanding growth behavior in clonal cell culture systems (section 2.1.4.1), obtaining 'real' single-cell information from pheno-seq data without profiling single cells would be of high importance to distinguish between genes that are generally associated with spheroid phenotypes and those who are robustly expressed at the single cell level. Therefore, we aimed to computationally infer single-cell regulatory states by deconvolution of pheno-seq data using both image information and a maximum likelihood inference approach. First, we generated a high-resolution imaging reference dataset from spheroids of different sizes by 3D light-sheet microscopy, which we used to determine the relationship of spheroid size and nuclei counts in order to estimate cell numbers from CRC spheroid pheno-seq imaging data (Figure 2.11a). As the PAGODA-normalized pheno-seq data exhibited no correlation between detected genes and estimated cell numbers (Figure 2.11b), we downsampled the data to achieve a uniform number of mRNA counts per estimated single cell content (Figure 2.11c).

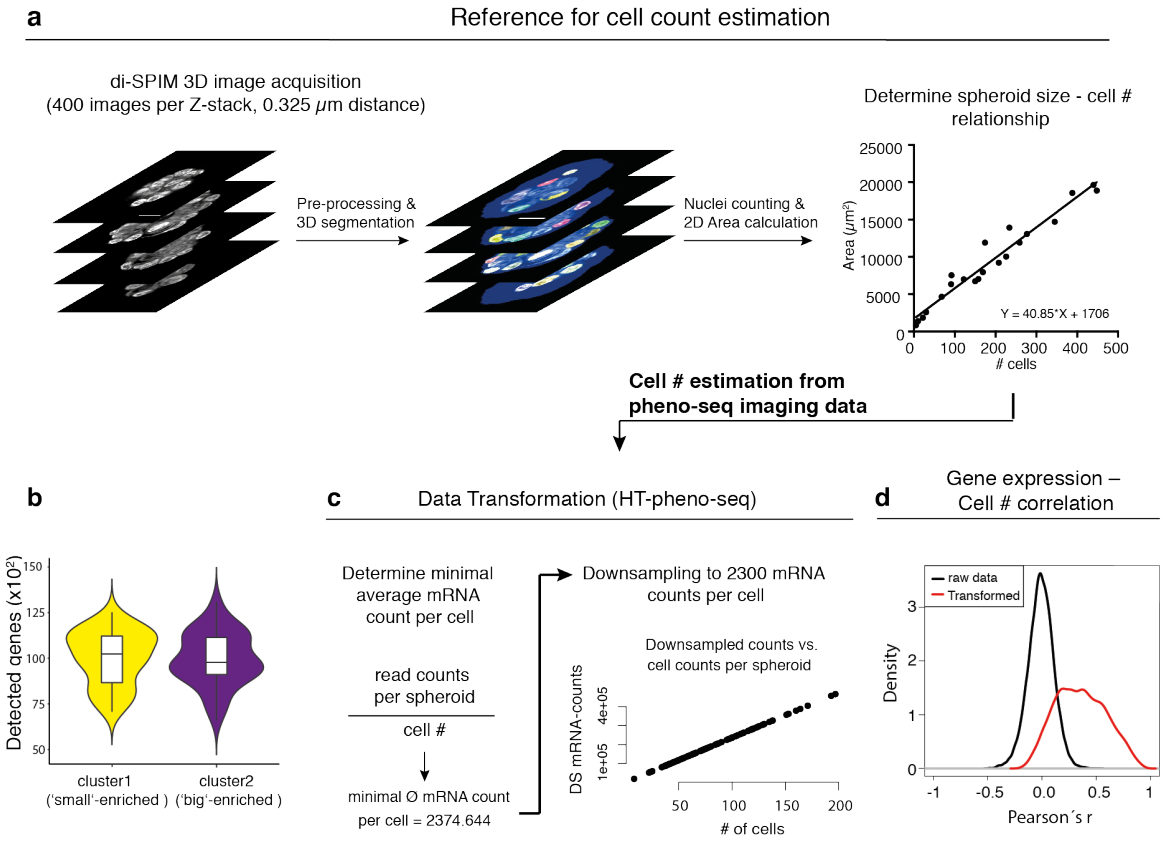


Figure 2.11 | Estimation of cell numbers from pheno-seq data and normalization by downsampling to estimated counts per cell. (a) Generation of a two color (Hoechst and CellTracker Red) 3D image reference dataset (20 spheroids) by using dual-view inverted selective plane microscopy (di-SPIM). 3D Segmentation and image analysis enables counting of nuclei. The calculated cell number – spheroid size relationship is used to estimate cell numbers from HT-pheno-seq data. **(b)** Violin plot showing detected genes plotted per cluster shown in Figure 2.7. Violin-plot center-line: median; box limits: first and third quartile; whiskers: ±1.5 IQR. **(c)** Correction for lost correlation of cell numbers and library complexity by transforming counts to approximated total mRNA abundances in spheroids of different sizes. Raw mRNA counts are divided by estimated cell numbers and the calculated minimal average mRNA count is used to normalize the data by downsampling counts to 2300 counts per cell in the CRC phenoSeq dataset. This strategy results in a perfect correlation of cell numbers and mRNA counts. The estimated cell number is plotted against normalized mRNA counts. **(d)** Pearson's correlation coefficients (r) distributions of gene expression and cell numbers for all 13,868 genes before and after data transformation. Light-sheet microscopy and associated image analysis in (a) has been done together with Björn Eismann and Friedrich Preußer. RNA-seq analysis in (b) has been jointly performed with Jeongbin Park. RNA-seq analysis in (c) and (d) has been jointly performed with Christiane Fuchs and Lisa Amrhein.

As expected, this transformation introduces a positive overall shift of correlations between gene expression and cell numbers compared to raw mRNA counts (Figure 2.11d), which can be mainly explained by housekeeping genes with a constant number of mRNA molecules per cell (Supplementary Figure 7a). However, heterogeneously expressed genes such as previously identified secretory markers TFF3 and DEFA5 do not exhibit any correlation with cell numbers (Supplementary Figure 7b and c), thereby validating our normalization approach.

To identify genes with heterogeneous single-cell regulatory states, we used a maximum likelihood inference approach previously developed to deconvolve cell-to-cell heterogeneities from random 10-cell samples[230] (Figure 2.12a). The adapted algorithm uses estimated cell numbers per spheroid to fit two log-normal distributions (LN-LN model) to given 'mixed-n' datasets in order to identify genes with bimodal expression pattern at the single-cell level (Stochastic Profiling, see Methods). Importantly, this approach unbiasedly identifies genes that are likely to show a heterogeneous and robust expression within spheroids at the single-cell level, instead of comparing gene expression between spheroids.

Whilst the deconvolution algorithm assumes that cellular subtypes are identically distributed across spheroid samples, pheno-seq is principally based on clonal spheroids whose cell number, subtype composition and transcriptional profile is dependent on the state of the founding cell. Based on the CSC model and the above indicated CRC differentiation hierarchy, we assume that continuously growing spheroids ('big' phenotype) harbor all cellular subtypes present in this culture system, including stem-like cells, whereas small spheroids with limited proliferative capacity and low cell numbers are more homogeneous and contain only differentiated subtypes. Thus, inferred single cell regulatory states should be enriched for genes specific for the stem-like compartment, as these represent the major source of heterogeneity at the single-cell level.

Deconvolution of the CRC pheno-seq dataset (n=95 spheroids) revealed 1,012 genes that exhibit an improved two-population fit as compared to a one-population fit, assessed by the Bayesian information criterion (BIC) to calculate the quality of the fit relative to the number of inferred parameters (Figure 2.12b). Most fits resulted in a highly-expressing cellular fraction of 5-15% (Figure 2.12c) thereby matching the fraction of cells with spheroid forming capacity in this model[204]. Interestingly, the positive shift of correlations between mRNA counts and cell numbers (before and after downsampling) is much more pronounced in two-population genes compared to non-two-population genes (Supplementary Fig. 13d), suggesting that many of the inferred two-population genes are involved in proliferative capacity. Indeed, gene set enrichment analysis revealed a high proportion of MYC targets and genes involved in the regulation of cell growth and proliferation (Figure 2.12d). In addition, high enrichment of genes involved in oxidative phosphorylation (OXPHOS) indicated for heterogeneous mitochondrial activity at the single-cell level, a phenomenon recently identified for intestinal stem cells and neighboring Paneth cells in the small intestine[192]. Strikingly, a high number of identified genes are overlapping with a recently identified stem cell signature of the small intestine revealed by massively parallel scRNA-seq[88], including SMOC2, APP, PRMT1, RGMB, MAPK1 and CTNND1, respectively (Figure 2.12e).

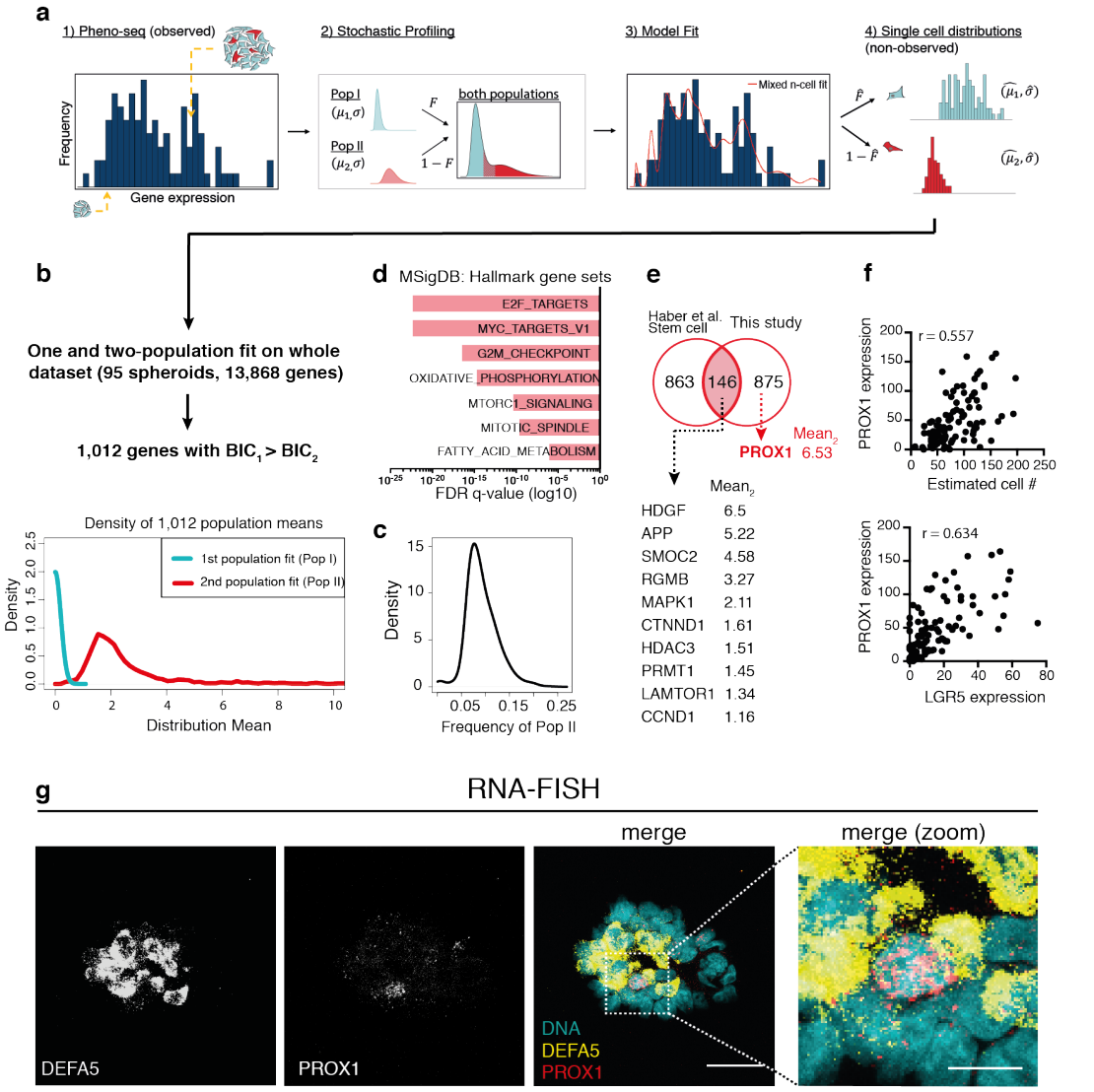


Figure 2.12 I Single-cell deconvolution of CRC spheroid pheno-seq data by maximum likelihood inference. (a) Adapted maximum likelihood approach[230] based on estimated cell numbers and transformed pheno-seq data (n = 95 spheroids): 1) Transformed pheno-seq data build a distribution of measurements for inference by the model. Coloring of cells in spheroids: red = stem-like; cyan = differentiated. 2) Assumptions on single cell distributions: Model of heterogeneous gene regulation in which individual cells are supposed to exhibit gene expression at low (Pop I) or high (Pop II) levels with common coefficient of variation. Four parameters of the model are the log-mean expression for each subpopulation (μ1 and μ2), the proportion of cells in the high subpopulation (F), and the common logarithmic standard deviation of gene expression (σ). 3) Based on the model in (2), a likelihood function is derived that takes different numbers of cells per spheroid into account. The likelihood function is maximized by searching through the four parameters of the model to identify those that are most likely given the experimental observations. 4) The four parameters define inferred single cell distributions of low and high-level populations. **(b)** 1,012 genes show improved two-population fit as compared to a one population fit (BIC: Bayesian information criterion). Densities of the means of the first (Pop I: low regulatory state) and second population (Pop II: high regulatory state) for all identified 1,012 genes. **(c)** Frequency distribution of single cells with high regulatory state (Pop II) of identified 1,012 genes. **(d)** Gene set enrichment analysis for two-population genes (1,012 genes) based on HALLMARK gene sets[231] derived from the MSigDB. Bar plot shows top enriched gene sets ranked by FDR q-values. **(e)** Overlap between identified two-population genes and murine intestinal stem cell signature from scRNA-seq study[88] shown as Venn-diagram. Selected genes are listed below ordered by mean for high-state population Pop II (Mean2). **(f)** Scatter plots for relations of PROX1 expression and estimated cell numbers (upper) and between PROX1 expression and expression of the major intestinal stem cell marker LGR5 (lower) as well as associated Pearson's correlation coefficients (r). **(g)** Validation by quantitative RNA-FISH. Co-staining of CRC spheroids by probes labelling PROX1 (Atto550) and DEFA5 (Alexa488) mRNA and Hoechst counterstaining for visualization of DNA. Merged images: DNA: cyan; DEFA5 yellow; PROX1: red. Images show Z-projections (scale bar 30 μm and 10 μm for magnified merged image). RNA-seq analysis in (a-f) has been performed together with Christiane Fuchs and Lisa Amrhein.

Here, we identified the transcriptional regulator PROX1 as gene with a high population mean (Pop II) (Figure 2.12e) that shows a strong correlation with cell numbers and also with expression of the major stem cell marker LGR5 (Figure 2.12f). In addition, PROX1 top correlated genes exhibited a strong overlap with the signature defining big spheres when relative gene expression differences between spheroids were analyzed (Figure 2.8). In the normal intestinal epithelium, PROX1 is specifically expressed in the enteroendocrine lineage[232]. However, two studies based on mouse tumor models suggest a role for PROX1 in CSC maintenance and metastatic outgrowth[233,234]. In line with these observations, we validated PROX1$^+$ cells as rare subpopulation in a patient-derived human tumor model by RNA-FISH (Figure 2.12g). Furthermore, PROX1$^+$ cells seemed to be framed by DEFA5$^+$ Paneth-like cells, suggesting a similar niche dependency for normal stem cells and CRC stem-like cells at distant sites of neoplasia. Taken together, gene expression deconvolution of pheno-seq data provides information about gene expression patterns at the single cell level even without acquiring additional single cell expression profiles.

2.2 Heterogeneous metabolic signatures are linked to cancer cell differentiation in a 3D model of colorectal cancer

2.2.1 scRNA-seq of 12 spheroid lines derived from CRC patients

2.2.1.1 Culture of CRC spheroid cultures with unique sets of driver mutations

Colorectal cancer (CRC) is the third most common cancer worldwide, causing approximately 10% of all cancers[235]. On the molecular level, CRC tumors have been classified in detail by bulk DNA and RNA-seq in order to identify markers or marker profiles with prognostic and therapeutically predictive value[161,236]. In general, CRC tumors can be subdivided into hypermutated (16% of cases) and non-hypermutated (84% of cases) tumors. Hypermutation is caused by mutations in DNA-mismatch repair genes that lead to the accumulation of DNA mutations mainly in repetitive microsatellite fragments (microsatellite instable, MSI). In contrast, non-hypermutated tumors are typically characterized by chromosomal instability[236]. Several investigations could reveal critical genes and pathways that are relevant for the initiation and progression of CRC, of which the tumor suppressor and negative WNT regulator APC belongs to the most mutated genes in both hypermutated and non-hypermutated tumors. Importantly, nearly all CRC tumors carry activating mutations in the WNT pathway[236] which is known to maintain the undifferentiated state of stem cells at the base of the intestinal crypt (see section 1.3.1.2). On the gene expression level, CRC tumors can be subdivided into four consensus molecular subtypes (CMS) with defining characteristics related to immune infiltration (CMS1), WNT activation (CMS2), metabolic deregulation (CMS3) and stromal infiltration (CMS4) that largely coincide with MSI status and chromosomal instability[161]. In addition, nearly all CRC tumors seem to exhibit changes in targets of the transcriptional regulator MYC[236], a WNT downstream target that controls proliferation and differentiation in the normal intestine[162,237], suggesting an important role for MYC in the progression of CRC.

Despite differences between tumors in mutational signatures and bulk expression profiles, recent investigations could reveal a general hierarchical organization of CRC cells that is similar to the healthy intestinal crypt, including LGR5$^+$ stem-like cells as putative tumor cell of origin[238,239] that give rise to more differentiated cells with reduced tumorigenic potential[205] (see section 1.3.1.2). Relatively little is known about the tumor cell composition in CRC. For example, by using multiplexed single cell qPCR, Dalerba et al. could show that CRC tissue contains distinct cancer cell populations whose transcriptional profiles are similar to known intestinal lineages, including LGR5$^+$ stem-like cells and KRT20$^+$ differentiated cells[227]. However, the number of profiled genes is limited in this approach and no study has yet

applied unbiased scRNA-seq to dissect the molecular heterogeneity and cellular composition of tumor cells in CRC.

To fill this gap at least for 3D *in vitro* models, we aimed to dissect 12 floating spheroid lines derived from CRC patients by scRNA-seq. The cultures originate from different primary tumor sites and metastases, and include the culture derived from a liver metastasis that was used for pheno-seq in the previous chapter (Supplementary Table 2). As these 3D cultures lack stromal and immune cell types but reproduce functional CRC tumor cell heterogeneity[204] and maintain subclonal composition[226], they represent appropriate models of CRC complexity with a much lower number of single cells required to understand heterogeneity between tumor cells. Overall, we selected 10 lines with a unique set of driver mutations inferred by whole exome sequencing, including two spheroid cultures derived from MSI tumors, and two cultures with unknown genotype (Table 1).

Table 1 | Driver mutations and microsatellite status (MSI) of CRC cultures*

Patient #	ID	TP53	APC	KRAS	TTN	SOX9	SMAD4	FBXW7	PIK3CA	CTNNB1	TCERG1	TCF7L2	SMAD2	GPC6	MSH6	MYO1B	BRAF		MSI
P1		■	■	■	■		■												
P2		■	■	■			■												
P3		■	■	■		■													
P4		■	■	■	■	■	■		■	■	■	■	■		■				
P5						■							■						
P6		■																	
P7			■	■	■														
P8		■															■		
P9		■		■			■												
P10																			
P11																			
P12		■	■	■	■	■		■	■	■	■			■	■	■			

(*Whole exome sequencing information from Dr. Claudia Ball/Prof. Dr. Hanno Glimm)

2.2.1.2 Generation of single cell RNA sequencing libraries and classification of tumors

Depending on the growth rate, spheroids were grown for 6-14 days after trypsinization (Supplementary Table 3). In order to avoid secondary cell culture artifacts that might affect gene expression heterogeneity, including hypoxic cores in the inner regions of spheroids larger than 300-400 μm in diameter[240], we did not grow spheroids to sizes larger than 200 μm. Spheroid morphologies differed strongly between patients and most likely reflect grades of differentiation, ranging from compact spheres to loose and instable cell-cell connections.

Moreover, spheroid morphologies were correlated with the required dissociation time to generate single cell suspensions (Supplementary Table 3).

To obtain RNA-seq expression profiles of single cells, we used the TakaraBio iCELL8 platform to generate 3´-end sequencing libraries (see section 1.2.2.2). By using the integrated imaging system, we detected single cells by nuclear staining with Hoechst and excluded dead cells detected by propidium iodide. Notably, several cultures were difficult to dissociate despite applying high Trypsin concentrations and shear forces or exhibited strong tendencies to quickly rebuild cell clusters after dissociation. Therefore, we needed to exclude several cell multiplets manually that could not be detected by the image analysis software provided. After sequencing, pre-processing and library QC (see Methods 5.2.4.1), we obtained 4663 single cell profiles, an average of 389 cells per patient, and detected on average more than 4000 genes per cell (Table 2).

Table 2 | scRNA-seq library information of cells derived from CRC spheroids, LGR5 score and predicted consensus molecular subtype

Patient #	ID	Mean reads per cell	Mean detected genes (> 0 counts) per cell	Cell number after QC	LGR5 score*	CMS
P1	[redacted]	348,016	3535	325	12.85	3
P2		261,595	4072	309	0.23	-
P3		460,471	4537	551	6.43	2
P4		1,061,813	4186	263	87.72	3
P5		334,099	3943	502	4.61	-
P6		1,276,856	5116	141	0.03	3
P7		359,362	4335	434	10.18	-
P8		190,170	4174	197	3.38	4
P9		527,407	4354	464	0.00	4
P10		391,680	3418	736	3.35	-
P11		505,439	4036	308	1.43	3
P12		454,258	3977	433	0.00	3

(*Total LGR5 counts divided by cell number)

To approximate the abundance of CSCs in individual patient cultures, we scored each patient for the presence of reads that map to the CSC marker LGR5. Notably, the abundance of LGR5 reads per patient differed strongly, indicating different proportions of CSCs in individual patients (Table 2). Moreover, LGR5 gene expression may be globally deregulated in individual CRC samples. In support of this, four patients exhibited only very low numbers of LGR5 reads, indicating that these tumors are LGR5-negative[241]. Next, we classified individual cultures for their CMS using pseudo-bulk profiles that were assembled

from single cell expression counts[161,242]. Whereas eight cultures could be significantly classified as CMS 2,3 or 4, four cultures could not be assigned to any CMS, potentially because they belong to CMS1 which is normally characterized by immune infiltration.

2.2.1.3 Seurat scRNA-seq analysis reveals patient-specific clustering and gene expression

Next, we explored single cell expression profiles derived from all 12 patients using the Seurat analysis tool[115] (see Methods 5.2.4.2). 2D visualization using tSNE maps revealed that tumor cells cluster according to their tumor of origin (Figure 2.13a), a phenomenon that that has been observed for multiple tumor entities[87,168,169]. Hundreds of genes were preferentially expressed per individual tumor. Hierarchical clustering based on the top 10 differentially expressed genes per patient showed that tumor cells cluster, with one exception, by the tumor site they originate from but not by microsatellite status or CMS (Figure 2.13b). Tumor-specific top differentially expressed genes contained many WNT signaling components and downstream targets (e.g., FRZB, DKK1, TCF4, SOX2) as well as secretory differentiation markers (e.g. MUC12, MUC17, SPINK1, SPINK4, DEFA5, DEFA6), indicating that tumor-specific perturbations induce expression of different sets of signaling components and lineage-specific genes (Figure 2.13b).

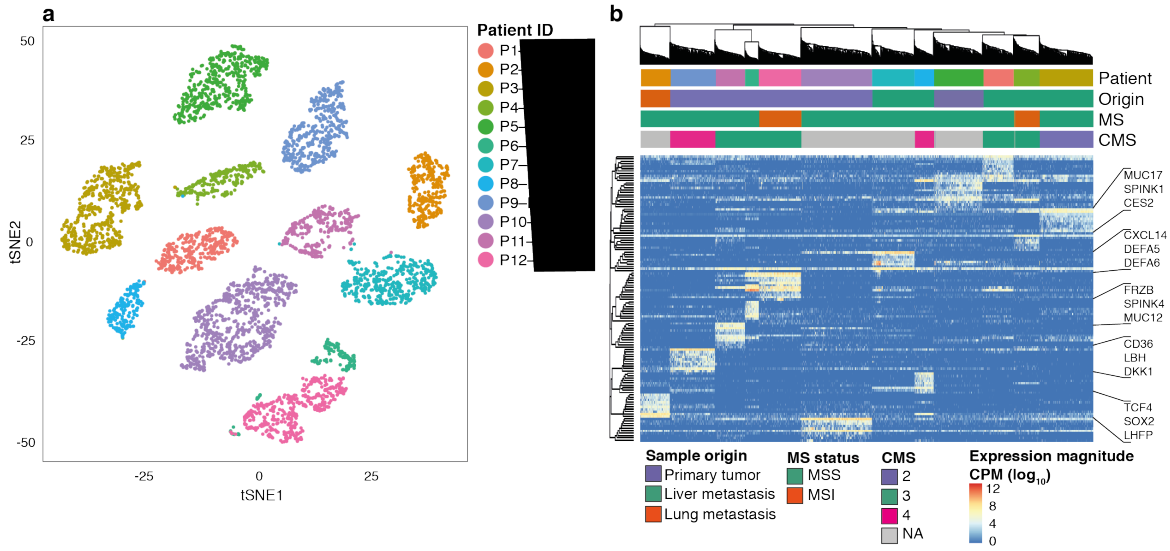


Figure 2.13 I 2D visualization and clustering of CRC single cell expression profiles reflects inter-patient variability. (a) 2D tSNE embedding of scRNA-seq profiles colored by patient. **(b)** Heatmap showing hierarchical clustering of absolute single cell gene expression based on top 10 differentially expressed genes per patient. Dendrograms reflect overall clustering and rows below show metadata information, including patient ID, sample origin (tumor site), microsatellite status (MS) and predicted consensus molecular subtype (CMS) of pseudo-bulk profiles per patient based on CMS-caller[242]. Three example genes for 5 selected patients are shown beside the heatmap. Data analysis was jointly performed with Teresa Krieger.

2.2.1.4 Analysis of relative single cell expression reveals shared metabolic heterogeneity

In order to identify shared expression programs across patients, we corrected for inter-patient variability in gene expression by calculating relative expression levels for each patient individually by mean-centering (see Methods 5.2.4.3)[87,89,243]. Upon correction, patient-specific clustering is completely eliminated, enabling the identification of potential gene expression programs that are shared across patients (Figure 2.14a). As classical clustering approaches did not result in distinct clusters, we analyzed variable gene expression by principal component analysis (PCA). Genes with high PC scores in the first principal component (PC1) were associated with hypoxia, Tumor Necrosis Factor alpha (TNF-α) signaling via NFκB and glycolysis, whereas genes with low PC scores are linked to cellular proliferation, growth and OXPHOS (Figure 2.14b).

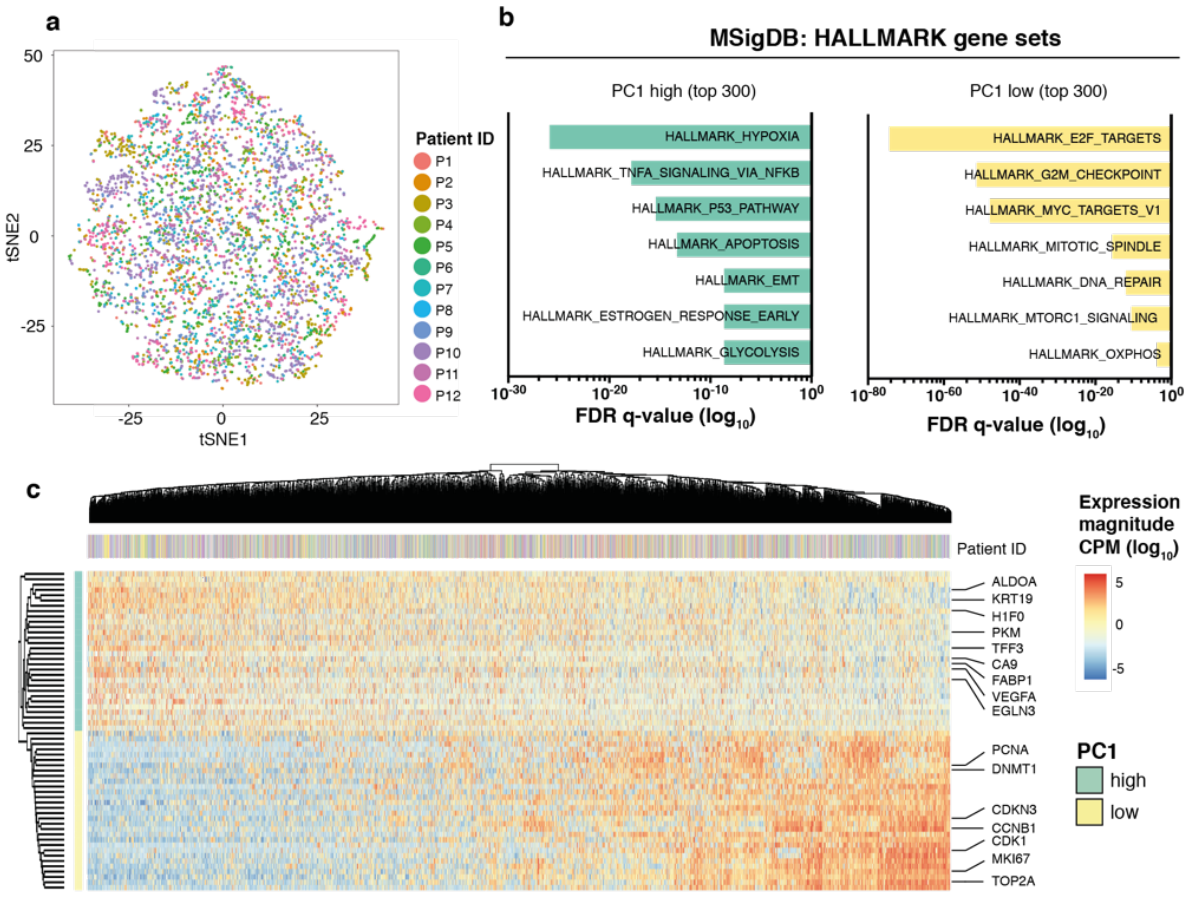


Figure 2.14 | Identification of intra-patient variability by mean-centering and PCA. (a) 2D tSNE visualization of mean-centered scRNA-seq profiles colored by patient. **(b)** Gene set enrichment analysis of top 300 genes with highest and lowest PC scores for first principle component using hallmark genes sets[231]. Bar plot shows top enriched gene sets ranked by FDR q-values. **(c)** Heatmap showing hierarchical clustering of top 30 genes with highest and lowest PC scores for first principle component of mean centered scRNA-seq data across all patients. Dendrograms reflect overall clustering and row below shows patient ID information. Example genes are listed beside the heatmap. Data analysis was jointly performed with Teresa Krieger.

Hierarchical clustering of the top 30 genes with high and low PC1 scores showed a clear anti-correlated pattern independent of patient origin. More detailed examination of the genes with highest PC1 scores revealed many intestinal differentiation markers (e.g., KRT19, TFF3, FABP1)[88] as well as the epigenetic regulator H1 Histone Family Member 0 (H1F0) that is primarily expressed in terminally differentiated cells[244]. Moreover, expression of H1F0 has been shown to be anti-correlated with self-renewal and tumorigenic capacity of cancer cells[245]. As these differentiation markers co-varied with hypoxic (e.g., CA9, EGLN3, VEGFA) and glycolytic markers (e.g., ALDOA, PKM), and cell cycle genes were correlated with OXPHOS genes, we reasoned that self-renewal and tumor cell differentiation are associated with metabolic preferences in CRC. Previously, OXPHOS has been linked to stem cell metabolism in normal[192] and cancer[199] intestinal tissues. However, we could neither identify known stem cell markers in the first PC, nor could we identify gene expression signatures in the other PCs that related to intestinal lineages. Thus, we reasoned that PCA is capable of revealing broad gene expression variability across patients but is limited in identifying gene expression programs that are expressed by a lower number of cells.

2.2.1.5 Non-negative matrix factorization identifies metabolic gene expression programs and signatures specific for intestinal cell types

In order to more accurately identify variable transcriptional programs across patients, we adapted a computational approach based on non-negative matrix factorization (NNMF)[246] that has been used to identify EMT related gene expression signatures in primary and metastatic head and neck cancer cells[169] (see Methods 5.2.4.3). The adapted NNMF workflow (Supplementary Figure 8) included the following steps: (i) inference of variable gene expression programs (factors) by NNMF, (ii) identification and removal of patient-specific factors (Supplementary Figure 9), (iii) evaluation of factors for biological relevance (Supplementary Figure 10), (iv) definition and clustering of core meta-signatures (Supplementary Figure 11), (v) computation and clustering of binary ON/OFF cell states per core meta-signature. We applied NNMF only to the eight LGR5^{+} spheroid lines (Table 2) in order to identify lineage-specific transcriptional programs including a potential CSC population. Of the 25 initial factors, we excluded eight that were preferentially expressed in individual patients. Next, we removed an additional four factors with high enrichments of genes involved in RNA binding and processing (e.g., GO_POLY_A_RNA_BINDING: FDR q-value 1.86 x 10^{-32}), as they are most likely associated with technical variation.

The remaining 13 factors could be classified into two main categories: those linked to known intestinal 'lineage' (or cell type) marker genes, and the rest which we defined as transient or

oscillatory 'cell states' (Supplementary Figure 10). The identified cell states can be further subdivided into three subcategories. The first contains three factors enriched for genes involved in the regulation of proliferation, including cell cycle stages G1/S (e.g., PCNA, RRM2, MCM4/10, BRCA2) and G2/M (e.g., TOP2A, CENPF, MKI67, CDK1) as well as MYC targets (e.g., CCND1, MYC, MINA). The second is characterized by genes involved in immune and stress responses, including antimicrobial chemokines (e.g., CXCL1/2/3, CCL20, IL-8), TNF-α signaling via NFκB (e.g., HBGEF, FOS, FOSL2, IL18), and Interferon signaling (e.g., ISG20, STAT1, STAT3, IRF7, IFI27). The third contained five factors with genes relating to diverse metabolic functions, including OXPHOS (e.g., PRDX3/4, ATP5O, ATP5G1, COX16), fatty acid metabolism (e.g., CES2, RETSAT, FABP2) and hypoxia/glycolysis (e.g., HILPDA, VEGFA, CA9, PGK1, CKB, ALDOA). One of the two factors enriched for genes involved in hypoxia/glycolysis contained many genes that were associated with high PC1 scores (Figure 2.14) and also included many differentiation markers (e.g., H1F0, TFF3, FABP1, KRT20, KRT19, MUC13). Thus, this factor overlapped with the second 'lineage' category, which also comprised a stem-like factor with many known markers for normal intestinal stem cells, including SMOC2, PTPRO, SP5, RGMB and LGR5[88,229]. This factor also contained the transcriptional regulator PROX1 that we identified as a putative CSC marker by pheno-seq (Figure 2.12). Moreover, NNMF enabled the identification of a Paneth-like subpopulation also detected by pheno-seq, which might serve as cellular niche for CSCs (e.g. DEFA5, DEFA6, FCGBP, MUC2)[88,229].

Next, we computed meta-signatures scores for each cell based on the averaged expression of the top 200 genes per factor and merged signatures that exhibit similar enrichments and clustering patterns (Supplementary Figure 10 and Supplementary Figure 11), resulting in eight 'core' meta-signatures. Hierarchical clustering of these core meta-signatures showed that signature gene expression is independent of patient origin, similar to PCA (Figure 2.15a). Although NNMF analysis did not result in mutually exclusive patterns of gene expression signatures, clear tendencies for discrete and overlapping transcriptional programs were visible. For example, the cell cycle, OXPHOS and MYC-target signatures exhibited a pronounced overlap. This observation indicates that high proliferation rates in this putative transit-amplifying (TA) compartment are driven by MYC and accompanied by OXPHOS, similar to hematopoietic progenitor cells[247–249]. In contrast, high signature scores for immune responses, hypoxia/glycolysis, stem and Paneth cells showed a relatively exclusive pattern.

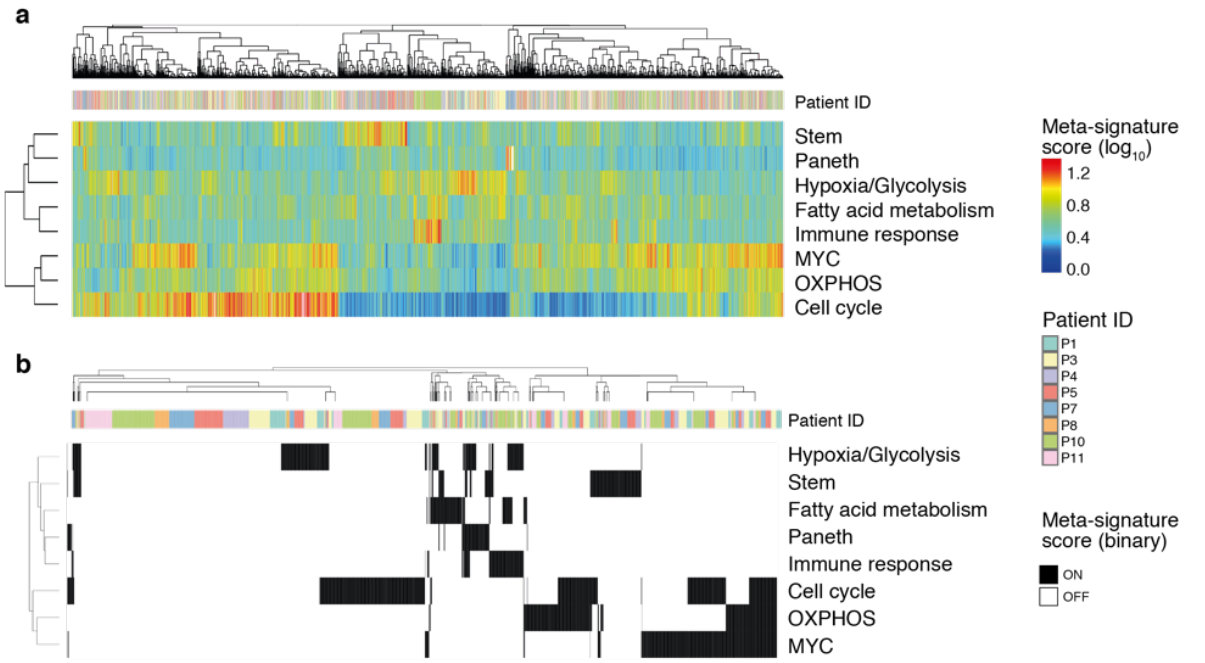


Figure 2.15 | Identification of shared gene expression programs in eight LGR5⁺ CRC patients by NNMF. **(a)** Heatmap showing hierarchical clustering of core-meta signature scores identified by NNMF across cells fromeight LGR5⁺ tumors. Dendrograms reflect overall clustering and row below shows patient ID information. Names of meta-signatures reflecting enriched genes are listed beside the heatmap. **(b)** Heatmap showing hierarchical clustering of binary core-meta signature scores (reflecting ON/OFF states of respective signatures per cell) across cells from eight LGR5⁺ tumors. Dendrograms reflect overall clustering and row below shows patient information. Names of meta-signatures reflecting enriched genes are listed beside the heatmap. For 948 cells, no 'ON' state could be assigned for any core meta-signatures. Data analysis was jointly performed with Teresa Krieger.

In order to assess whether core meta-signatures are active in individual cells, we inferred binary ON/OFF states of defined core meta-signatures per cell (see Methods 5.2.4.3), thereby generating a reduced representation of signature scores. Hierarchical clustering of binary states resulted in a similar pattern as for meta-signature scores but also revealed a partial overlap of fatty acid metabolism, Paneth signature, immune response and, to some extent, hypoxia/glycolysis (Figure 2.15b). Quantification of cellular fractions per patient revealed varying proportions of active core meta-signatures that can be linked to differential expression between patients (Supplementary Figure 12). For example, P4 had the highest LGR5 score and exhibited the highest fraction of cells with an active stem signature. Moreover, Paneth markers DEFA5 and DEFA6 were preferentially expressed in P7, which showed the highest number of cells with an active Paneth signature. Taken together, NNMF enabled the inference of lineage-specific transcriptional signatures and overlapping cell states across tumor cells of multiple CRC patients.

2.2.1.6 Lineage-specific metabolic preferences in CRC

Metabolic heterogeneity can be tightly linked to cellular differentiation in healthy and neoplastic tissues (see section 1.3.1.5). Rodriguez-Colman et al. recently revealed differential metabolic programs and associated interdependencies between LGR5^{+} stem cell cells (OXPHOS) and niche-forming Paneth cells (glycolysis) in the healthy small intestinal epithelium, but the metabolic state of other cell types was not determined[192]. In CRC, several indications suggest a dependency of CSCs on mitochondrial function[199] but associated genetic programs and metabolic preferences of other cancer cell subtypes in CRC are unknown.

Based on both PCA and NNMF analysis, variable metabolic programs seem to be a major source of heterogeneity across CRC patients, but lineage-specific metabolic states could only be partially inferred. Whereas glycolysis/hypoxia can be assigned to terminally differentiated (Tdiff) cells (FABP1^{+}/H1F0^{+}), OXPHOS strongly overlaps with the MYC target signature and actively cycling cells that might, at least in part, represent the TA compartment that is located above the crypt in the normal intestine[57,159] (Figure 2.15). From our binary 'ON'/'OFF' categorization of cells that was based on a comparison of gene expression across the entire data set (Supplementary Figure 13) no distinct metabolic preferences could be directly assigned to stem and potentially niche-forming Paneth meta-signatures as previously described for the healthy intestine[192]. However, we reasoned that differential metabolic trends between stem and Paneth cells could be masked by much higher or lower expression of individual metabolic signatures in highly cycling cells or the terminally differentiated subpopulation, which Rodriguez-Colman et al. did not report on. For example, it is known that few intestinal cells are actively cycling in crypts compared to TA cells[57]; consistently, in our binary classification, we observed significant overlap between the MYC signature, OXPHOS and cell cycle, but only a very small overlap between stem cells and the OXPHOS or cell cycle signatures. We therefore conducted a pair-wise comparison of cell state meta-signature expression across the identified intestinal lineages, and further refined our scoring approach by focusing solely on genes that are known to relate to the inferred metabolic identities of meta-signatures, including cell cycle, OXPHOS, hypoxia and glycolysis (see Methods 5.2.4.3).

As expected, the strongest differences in metabolic states were visible between the Tdiff (FABP1^{+}) and MYC^{+} subpopulations, showing that most MYC^{+} cells are actively cycling and OXPHOS^{high}, whereas Tdiff cells exhibit high hypoxia and glycolysis scores but low cell cycle and OXPHOS scores. Although the differences are less pronounced, we detected similar and highly significant trends for stem and Paneth cells. Compared to Paneth cells, stem cells

show increased OXPHOS and lower hypoxia/glycolysis scores and *vice versa* (Figure 2.16b, c, and d), in line with previous observations in the healthy intestine[192]. Interestingly, the stem-like signature is also associated with enhanced expression of OXR1 and PON2. These genes are essential for the protection against oxidative stress[250,251] and might thus represent a compensatory mechanism against high ROS levels due to enhanced rates of OXPHOS. Additionally, the stem-like signature contains the gene Glutamate Ammonia Ligase (GLUL) that catalyzes the synthesis of glutamine (a major source for OXPHOS[252]) as well as MAP2K6, an essential component of the p38 signaling cascade[253], whose activity is known to coincide with high OXPHOS levels in intestinal stem cells[192]. At the same time, we observed that cell cycle scores are only slightly higher in stem cells compared to Paneth cells, indicating that putative CSCs are slow-cycling in CRC. In sum, these results show that proliferation as well as tumor cell differentiation are linked to metabolic identity in this model of CRC, that also includes putative LGR5⁺ stem and DEFA5⁺ Paneth-like subpopulations.

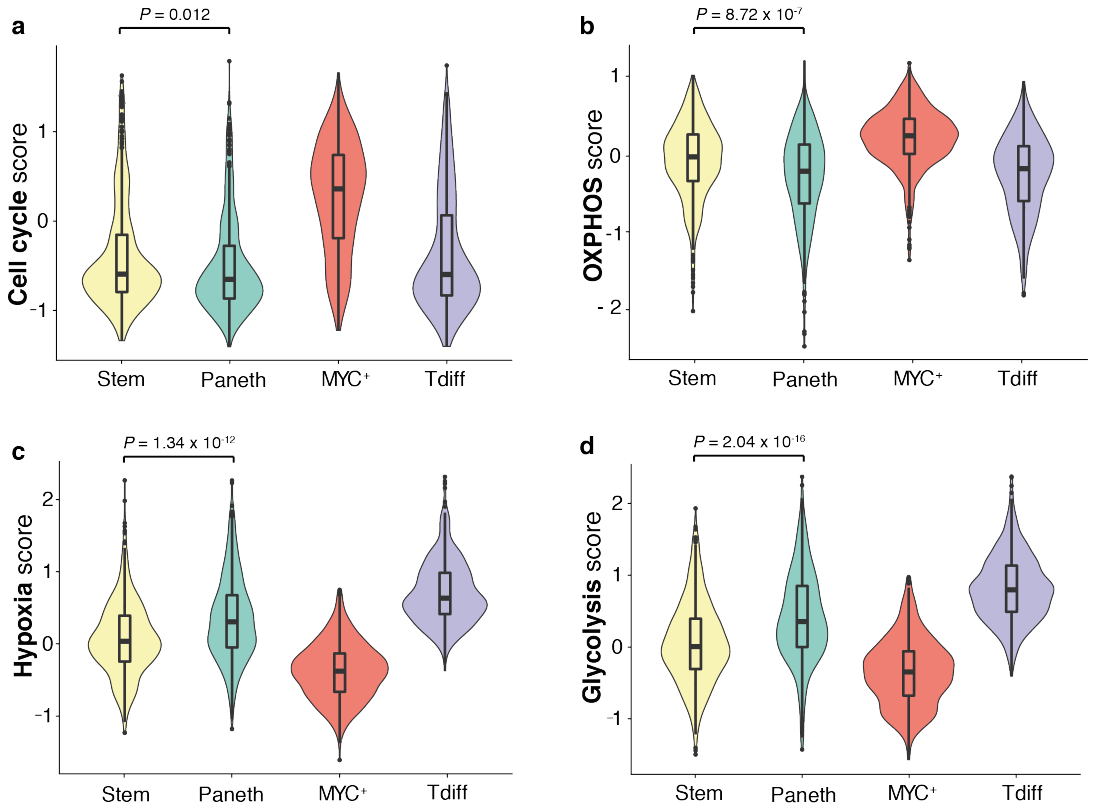


Figure 2.16 | Paneth and stem cell meta-signatures are associated with distinct metabolic tendencies.

Violin plots reflecting cell state scores (cell cycle, OXPHOS, glycolysis, hypoxia) for cells with active lineage-specific meta-signatures based on binary classification (see Figure 2.15b): Paneth (n=381), stem (n=458), MYC⁺ (n=552) and Tdiff (n=483). Violin-plot center-line: median; box limits: first and third quartile; whiskers: ±1.5 IQR. Indicated are *P*-values from unpaired two-tailed Students t-test.

2.2.2 In situ analysis reveals metabolic compartmentalization and potential cellular interdependencies

Despite the rich information content of scRNA-seq data, inference of spatial locations, morphologies and cellular interactions requires additional *in situ* analysis based on microscopy. Therefore, we aimed to map identified molecular markers to their spatial location in spheroids by using live-dyes and RNA-FISH[56]. We selected three representative CRC spheroid lines with distinct morphologies (Figure 2.17a), including one line derived from a primary tumor (P5) and two from liver metastases (P1 and P4).

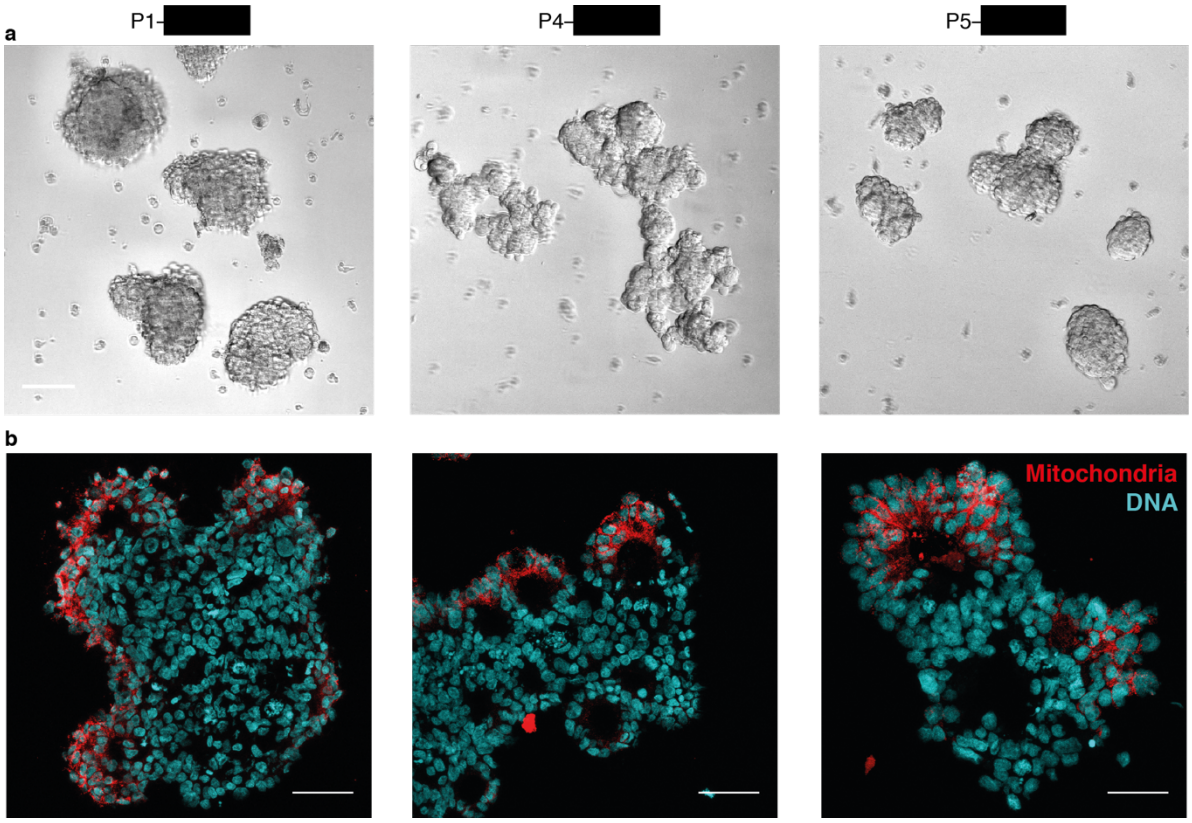


Figure 2.17 | Morphological heterogeneity and mitochondrial abundance in three CRC spheroid cultures.

(a) Brightfield images of three representative CRC spheroid cultures; Scale bar 100 μm. **(b)** Histological sections of CRC spheroid cultures shown in (a). Images represent maximum projections of 10 μm slices stained for nuclei (DAPI, Cyan) and mitochondria (MitoTracker Red CMXRos, Red); Scale bar 50 μm.

Histological examination revealed intestinal crypt-like phenotypes in all three spheroid lines, including luminal structures and partially polarized cells (Figure 2.17b), thus indicating for a high degree of differentiation in line with scRNA-seq results. To further validate scRNA-seq data, we stained spheroids with a mitochondrial live-dye (Mitotracker Red CMXRos) before histological preparation in order to asses metabolic heterogeneity for OXPHOS. In all three

patients, we could detect strong differences in mitochondrial content between cells and a compartmentalization of cells with high numbers of mitochondria especially at the outer layer of crypt-like regions (Figure 2.17b). Similar patterns have been observed in healthy intestinal organoids, where crypt formation ('budding') and associated differentiation are driven by OXPHOS and ROS signaling[192]. As the abundance of mitochondria is correlated with increased mitochondrial activity in intestinal organoids[192], we defined outer regions with high Mitotracker signal as OXPHOS^{high} regions.

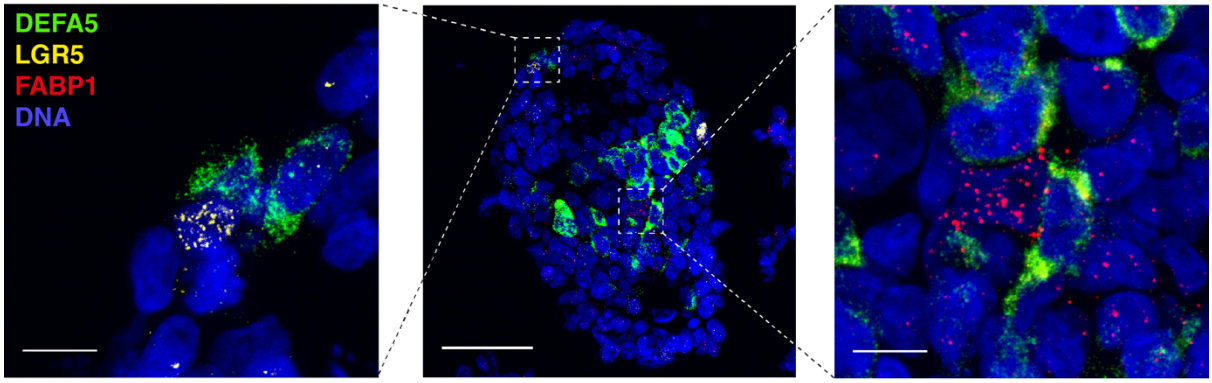


Figure 2.18 | Validation of lineage specific marker genes by RNA-FISH. Histological section of CRC spheroid derived from P1 co-stained for representative lineage-specific marker genes by RNA-FISH. Middle: overview image; scale bar 50 μm. Left and right: magnified images that represent dashed box regions in overview image (4x digital zoom); scale bar 10 μm. DEFA5: Paneth cells (green), LGR5: stem cells (yellow), FABP1: differentiated cells (red)**.** Images represent Z-projections from 10 μm slices. DNA counterstain by DAPI (blue). For P4 and P5 see Supplementary Figure 14.

Next, we used RNA-FISH to visualize mRNA abundance of subtype markers that we identified by scRNA-seq. Multiplexed RNA-FISH for representative intestinal lineage markers LGR5 (Stem), DEFA5 (Paneth) and FABP1 (TDiff) resulted in discrete staining of individual cells by either one or none of the three markers, thereby strongly indicating for the existence of distinct intestinal lineages in all three patients (Figure 2.18 and Supplementary Figure 14). Consistent with scRNA-seq results, spheroids from different patients differ both in abundance of subtypes (Supplementary Figure 12) and in expression magnitude of marker genes. In addition, clear tendencies of the spatial location of subtypes were visible. For example, DEFA5$^+$ cells primarily localized to the inner regions of spheroids, whereas LGR5$^+$ showed a tendency towards outer regions (Figure 2.18 and Supplementary Figure 14). In many cases, we could identify DEFA5$^+$ cells in direct proximity to LGR5$^+$ cells (Figure 2.18), as also observed for PROX1$^+$ cells (Figure 2.12), indicating for similar cellular interactions and dependencies between stem and Paneth cells as in the healthy intestine[4,229]. In healthy

intestinal organoids, crypt formation is associated with enhanced OXPHOS and with the emergence of Paneth cells and LGR5$^+$ stem cells at budding sites that later on develop distinct metabolic identities[192]. In CRC spheroids, we could detect enhanced mitochondrial abundance at budding sites (Figure 2.17), tendencies for the spatial location of subtypes (Figure 2.18 and Supplementary Figure 14) as well as anti-correlated trends in metabolic transcriptional programs for DEFA5$^+$ and LGR5$^+$ cells (Figure 2.16). Thus, we asked whether metabolic compartmentalization at budding sites is associated with the spatial distribution of lineage subtypes in CRC spheroids.

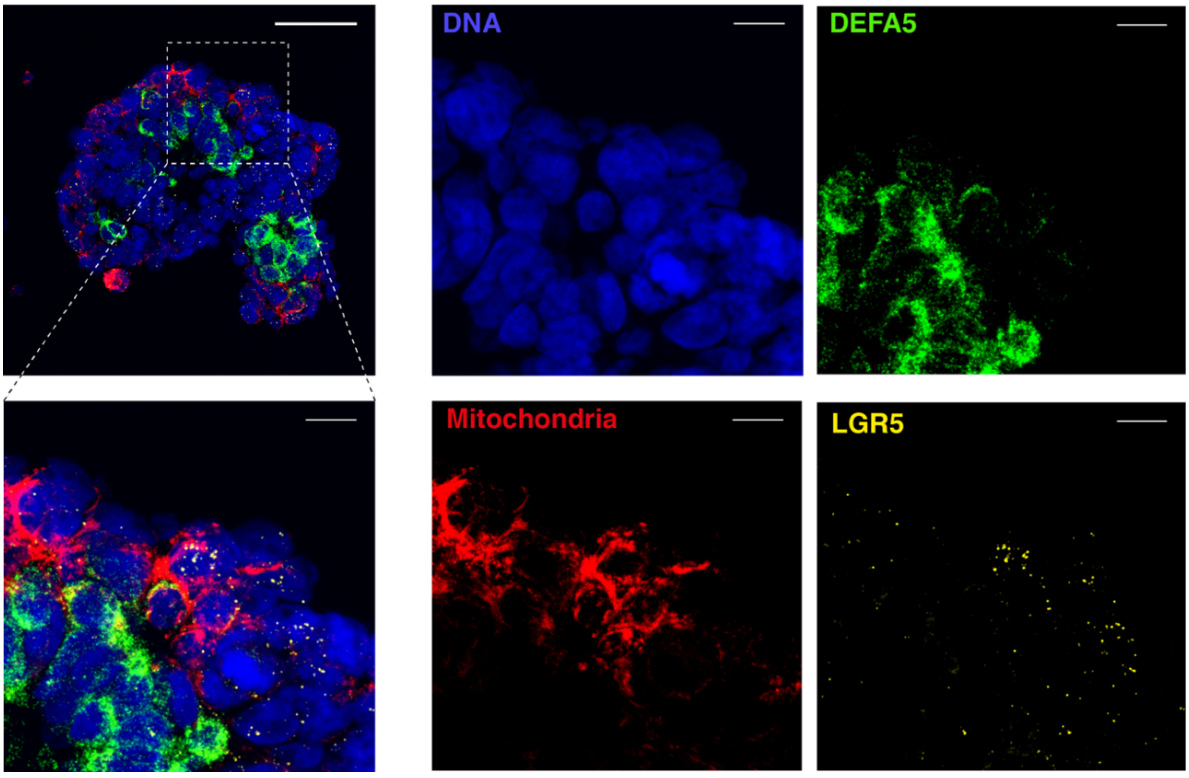


Figure 2.19 | The spatial location of LGR5$^+$ cells coincides with high mitochondrial abundance in budding spheroid regions. Histological section of CRC spheroid derived from P1 co-stained for representative lineage-specific marker genes by RNA-FISH and for mitochondria with Mitotracker Red CMXRos (100 nM). Top left: overview image (merge); scale bar 50 μm. Lower left: magnified image that represents dashed box regions in overview image (4x digital zoom); scale bar 10 μm. Middle and right: Individual channels of magnified image; scale bar 10 μm. DEFA5: Paneth cells (green), LGR5: stem cells (yellow), mitochondria (red)**.** Images represent Z-projections from 10 μm slices. DNA counterstain by DAPI (blue). For P4 and P5 see Supplementary Figure 15.

By combining mitochondrial staining and RNA-FISH, we could detect that DEFA5^{+} are largely excluded from OXPHOS^{high} regions. In contrast, LGR5^{+} cells were primarily located in OXPHOS^{high} regions in all three patients (Figure 2.19 and Supplementary Figure 15). However, LGR5^{+} cells are not restricted to the outer layer of CRC spheroids as we could also detect LGR5^{+} in inner regions of spheroids in some cases (Supplementary Figure 16). Differentiated FABP1^{+} cells do not locate to OXPHOS^{high} regions (Supplementary Figure 17), which is consistent with scRNA-seq results that revealed a co-expression of FABP1 with hypoxic and glycolytic markers (Figure 2.14 and Supplementary Figure 10). Furthermore, actively cycling cells stained by the cell cycle marker MKI67 were frequently but not solely present in OXPHOS regions in all patients (Supplementary Figure 18) which is in line with scRNA-seq data that showed a strong but not complete overlap of OXPHOS and cell cycle gene expression signatures (Figure 2.15 and Supplementary Figure 13).

In order to quantify observed compartmentalization of OXPHOS^{high} states and lineage-specific gene expression across multiple spheroids and thousands of cells, we developed an automated image analysis pipeline that includes (i) nuclei segmentation by deep learning, (ii) single cell quantification of binarized fluorescence signals, (iii) k-means clustering to identify 'ON' states in gene expression and mitochondrial abundance and (iv) computation of overlaps between 'ON' states of imaged channels (Figure 2.20 and Supplementary Figure 19). In line with scRNA-seq results and qualitative image evaluation, quantitative image analysis of overlaps between LGR5, DEFA5 and Mitotracker 'ON' states in three patients revealed a much higher fraction of LGR5^{+} cells that are OXPHOS^{high} at the same time compared to DEFA5^{+} cells (Figure 2.21). In addition, overlaps between LGR5^{+} and DEFA5^{+} cells were very low, showing that both markers define independent CRC subtypes.

Taken together, *in situ* analysis by microscopy validated scRNA-seq data and further indicated for lineage-specific metabolic preferences of putative stem and Paneth subtypes in CRC. Furthermore, metabolic tendencies of both lineage subtypes seemed to be associated with their spatial localization in spheroids, indicating for dynamic processes (e.g., cellular budding) that might influence functional tumor cell heterogeneity.

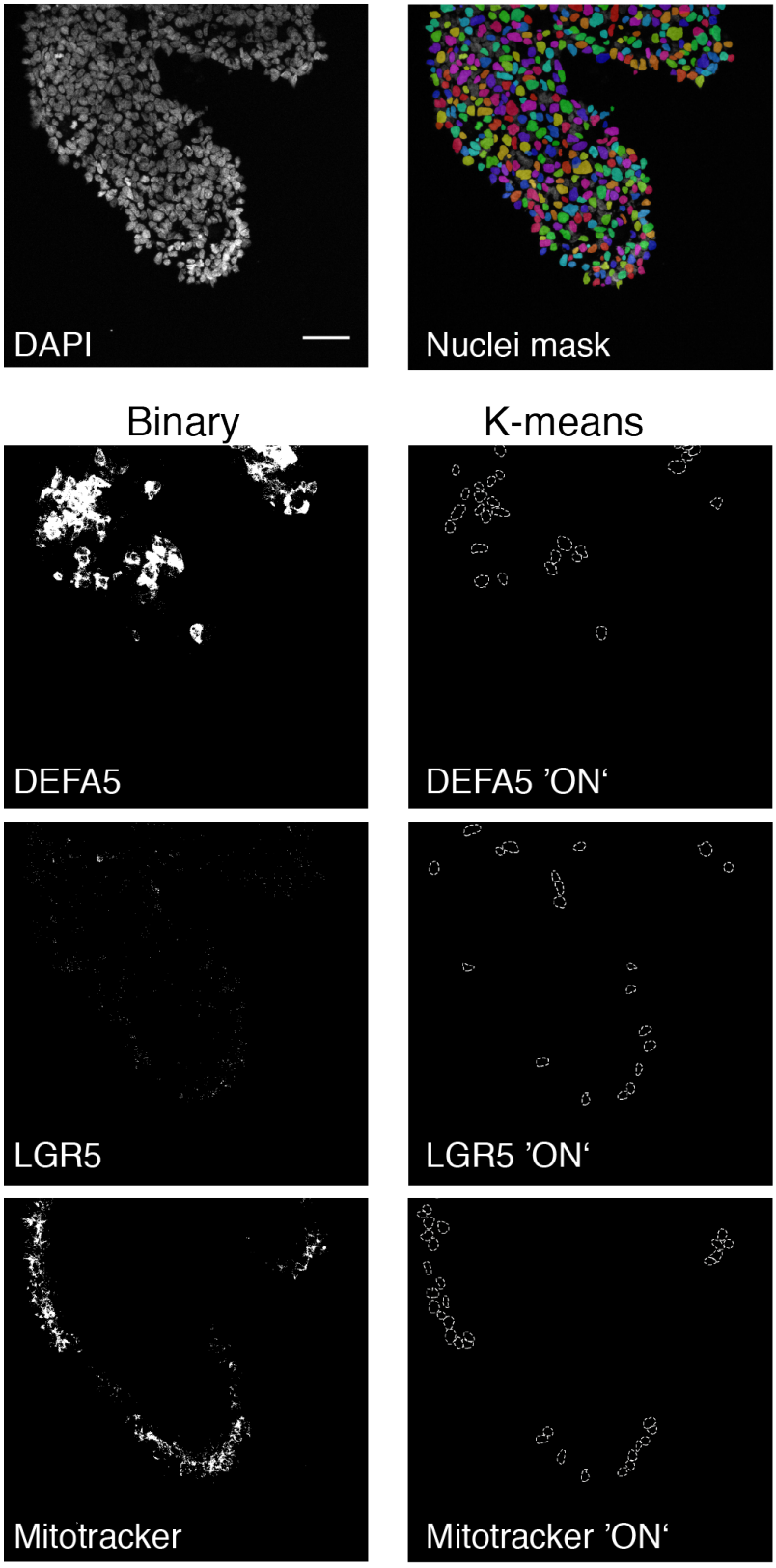


Figure 2.20 I Detection of single cell 'ON' states in gene expression and mitochondrial abundance by automated image analysis. Example images (same position/spheroid) from P1 reflecting major steps during image processing and analysis. Upper: raw DAPI signal (left) and nuclei detection by deep learning (right). Lower: Binarized fluorescence signals from RNA FISH probes (DEFA5 and LGR5) and Mitotracker Red (left) and detected single cell 'ON' states (white dashed circles) inferred by k-means clustering (right). Scale bar 50 μm. The image analysis pipeline was developed together with Foo Wei Ten.

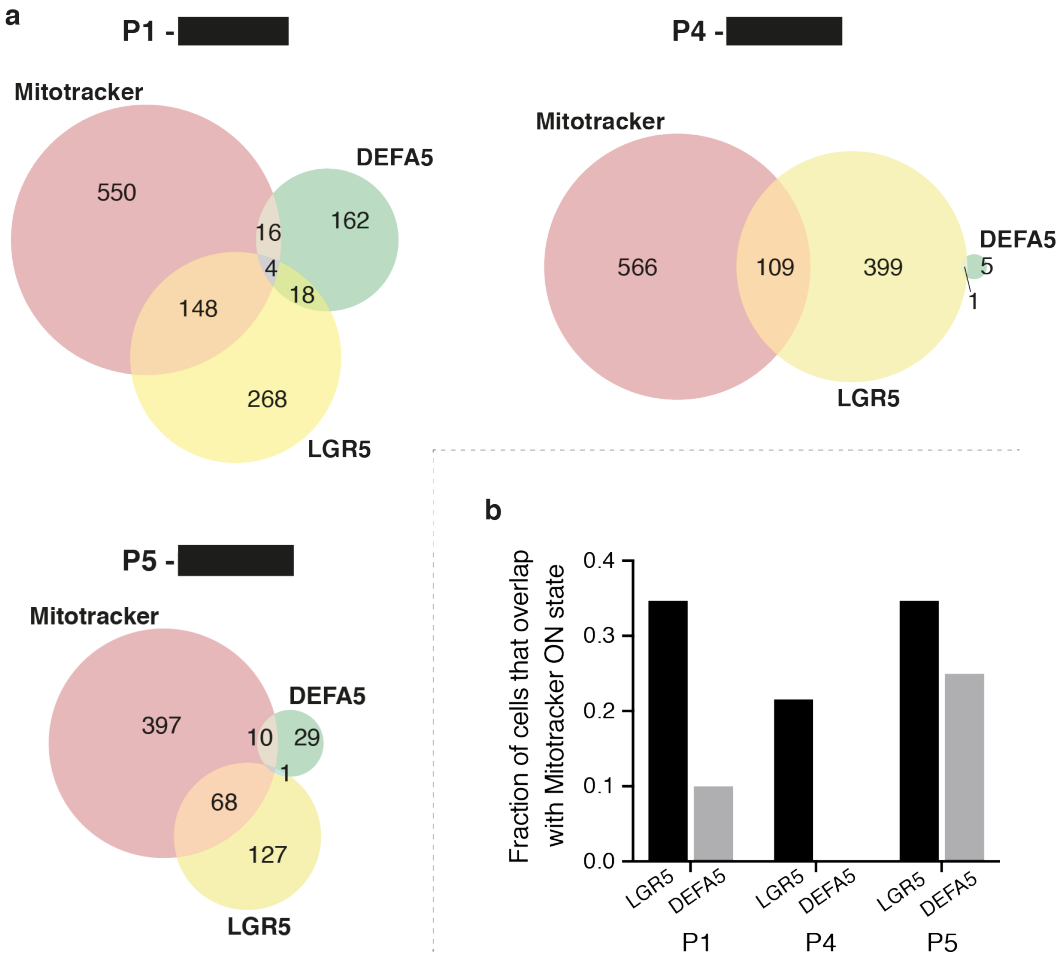


Figure 2.21 | LGR5⁺ cells show higher overlap with Mitotracker 'ON' states compared to DEFA5⁺ cells. (a) Venn diagrams showing overlaps between 'ON' states in gene expression (LGR5 and DEFA5) and mitochondrial abundance (Mitotracker Red) in three patients inferred by fluorescence microscopy and image analysis. Total number of analyzed cells: P1 n=4032; P2: n=3813; P3: n=2861. **(b)** Barplot reflecting fraction of LGR5⁺ and DEFA5⁺ cells that overlap with Mitotracker 'ON' states in three patients. The image analysis pipeline was developed together with Foo Wei Ten.

3 Discussion and Outlook

With this study, we provide a methodological framework to analyze tumor cell heterogeneity in 3D cell culture systems using combinations of NGS and microscopy. First, we developed pheno-seq as straightforward strategy for combined gene expression profiling and imaging of clonal tumor spheroids that enables a direct correlation of patho-phenotypes and underlying trancriptomes. Second, we applied scRNA-seq to 12 patient-derived CRC spheroid cultures and complemented results by RNA-FISH analysis *in situ*. Thereby, we could reveal a link between metabolic preferences and cancer cell differentiation, spatial spheroid organization and potential cellular interdependencies.

In the following, I will discuss the obtained results for both methodological strategies in the context of related work and previous studies. Furthermore, I will provide an outlook on required future work, potential applications as well as improvements regarding both technology and data analysis. In the final conclusion, I will relate both strategies to each other in the context of 3D cell culture systems as tool for personalized medicine. (Parts of section 3.1 have been adapted from the original publication for which I have written the original text[208])

3.1 Pheno-seq – linking morphological and functional features to gene expression in 3D cell culture systems

3.1.1 Pheno-seq – a complementary method to understand tumor cell heterogeneity

Imaging-based classification is used as standard method to classify tumor subtypes and disease states in primary samples, but this has not been broadly transferred to 3D cell culture systems, and if it was, only to compare inter-patient differences[211] rather than intratumor heterogeneity. On the other hand, recently developed technologies for scRNA-seq have been primarily used to dissect tumor cell heterogeneity in primary samples of several entities[87,89,168,169,254]. However, scRNA-seq does not provide a direct link to contextual cellular phenotypes since the available protocols involve dissociation of cells and loss of their multicellular context.

Spatial transcriptomics (see section 1.2.3) represents a promising approach to resolve intratumor heterogeneity both morphologically and transcriptionally[133,134]. Although this strategy would be of high relevance for 3D cell culture systems, *in vitro* cultures are not yet able to spatially reconstruct whole tumor mass, but rather small parts (spheroids, organoids) with characteristic morphologies and cellular subtype composition. Nevertheless, their

flexibility enables a much deeper analysis with various molecular tools and imaging technology (Figure 1.1), including time-resolved single cell analysis, which can inform about functional tumor cell heterogeneity and behavior.

Here we present pheno-seq, that complementary to spatial transcriptomics in primary samples, links cellular morphologies to underlying gene expression *in vitro* by directly combining high-throughput imaging and next generation sequencing of clonal tumor spheroids (Morpho-Transcriptomics). Besides functionally explaining heterogeneous behavior in 3D cell culture systems, pheno-seq also complements scRNA-seq in revealing heterogeneous gene expression by providing more RNA material than single cells, thereby increasing the gene detection rate per sample (Figure 2.2h) and decreasing drop-out rates. Despite lower cellular resolution than scRNA-seq, we show that pheno-seq is able to link cell type-specific genes to heterogeneous growth phenotypes even in a highly complex and patient-derived CRC cell culture system (Figure 2.7).

We could show that Paneth-like cells, are present in both small and big spheres, but with varying cellular ratios (Figure 2.8 and Figure 2.10). It has been shown previously that the fraction of functionally different subpopulations in this CRC model remains stable over several rounds of replating, suggesting that the composition of cells in continuously growing spheroids remains relatively stable[204]. Consequently, for a cellular subtype with limited proliferative potential in the putative CRC differentiation hierarchy, we would have expected a similar association of relative expression and size as observed for the $TFF3^+$ secretory signature above (Figure 2.7d). However, as we also detected several big spheres with high numbers of $DEFA5^+$ cells, we reasoned that Paneth-like cells exhibit a heterogeneous proliferative phenotype (high- and low-cycling) that could relate to the delayed-contributing subpopulation in CRC that has been described previously[204]. A recent study could show that, upon injury, proliferation and a stem-like transcriptomic profile is induced in a subset of Paneth cells that are post-mitotic under normal conditions[255]. Interestingly, this fate change depends on Notch signaling, which we identified as crucial for long-term proliferative potential of CRC spheroids (Figure 2.7 and Figure 2.9). Future studies will need to address the question whether a permanent injury or immune response[256] induces the observed proliferative phenotype of Paneth-like cells and analyze its functional implication in CRC development and progression.

Furthermore, we show that deconvolution by maximum likelihood inference provides an additional layer of information by revealing single-cell regulatory states of which many are likely to be associated with a distinct stem-like population (Figure 2.12), thereby further supporting a differentiation-like hierarchy in CRC. Based on our results, future studies should

shed light on functional characteristics and dependencies of the intestinal CSC compartment, potential cancer cell plasticity and the impact of subtype-specific metabolic preferences (further discussion on intratumor heterogeneity in CRC in section 3.2).

3.1.2 Future applications of pheno-seq

Pathologically relevant phenotypes are readily apparent from microscopy in both primary samples and in 3D culture systems. However, extracting functionally relevant features from both bulk and single cell RNA-seq data alone is difficult due to the high complexity of the data. Thus, adding visual information to Omics profiling provides an important functional layer of information.

Similar to spatial transcriptomics (section 1.2.3) and topographic single cell sequencing (section 1.3.1.1), we could identify molecular signatures that are associated with visual oncogenic phenotypes in 3D cell culture systems. These included invasive phenotypes that can be both identified *in* and *ex vivo*, as well as heterogeneous proliferative behavior whose characterization is restricted to *in vitro* systems. Thus, if 3D cell culture systems will be further embedded in personalized translational and clinical pipelines, pheno-seq might be a promising tool for identifying potential drugs targets to inhibit observed oncogenic behavior. For instance, we could show that perturbing identified CSC signaling with a γ-secretase inhibitor had a concentration-dependent negative effect on CRC spheroid growth (Figure 2.9). If the inhibitory effect is based on driving CSC into differentiation as described for mouse colonic stem cells[229] needs to be further validated.

An additional promising application for pheno-seq could be the mechanistic characterization of invasive cancer cell behavior. Defining the expression changes that are associated with visual phenotypes of cancer cell invasion is challenging due to high intratumor heterogeneity and technical limitations in isolating and characterizing cellular subpopulations. Hence, 3D cell culture systems and pheno-seq might be promising tools to analyze cancer cell motility over time and under defined conditions. As proof-of-concept, we could link EMT related gene expression programs to invasive cancer cell behavior in the MCF10CA breast cancer model, although it is not clear whether genetic or epigenetic alterations induce this phenotypic heterogeneity. In addition, the applicability of patient-derived material to characterize mesenchymal tumor cell phenotypes with pheno-seq still needs to be demonstrated.

Taken together, we expect that pheno-seq as combination of functional single cell growth assay with combined image and gene expression profiling will be widely applied in cancer biology, ranging from primary[34,35] to circulating tumor cells (CTCs)[257]. Moreover, pheno-seq might be extended to developmental biology based on model organisms where embryonic

development following fertilization is often not completely synchronized. Hence, pheno-seq might be used to resolve transcriptional changes that are associated with morphological transitions in non-synchronized developmental processes. Finally, pheno-seq might be applied in combination with single cell pooled-screening approaches[258] but with single cell growth and spheroid phenotype readout in addition to gene expression profiling (Figure 3.1).

3.1.3 Limitations and possible improvements of pheno-seq

3.1.3.1 Spheroid isolation

Pheno-seq represents a straightforward methodological strategy that does not necessarily require complex experimental setups (manual pheno-seq) and which can be applied to virtually any 3D culture system given that the phenotypic identity is maintained upon spheroid isolation. The isolation of spheroids or organoids from Matrigel might influence fragile phenotypes including highly invasive cells with long protrusions which should be kept in mind for downstream analysis and data interpretation. Fixation prior to isolation or alternative experimental setups where spheroids are imaged in Matrigel before isolation/lysis might be considered if the phenotype of interest cannot be maintained after isolation. However, the latter strategy would not be possible with the iCELL8 system.

3.1.3.2 Lysis, RT and cDNA amplification chemistry

Whereas lysis, RT and cDNA amplification reagents and reaction conditions for scRNA-seq have been highly optimized during the last years, there is still room for improvement for HT-pheno-seq. For instance, the gene detection rate is significantly higher for manual compared to HT-pheno-seq (Figure 2.2). As a consequence, low-expressed genes like the EMT master regulator SNAI2 could be identified as 'aberrant'-specific by manual pheno-seq, but not with the high-throughput approach. Although there is always a trade-off between throughput and sensitivity even for scRNA-seq[78], several steps of the HT-pheno-seq workflow might influence library quality compared to manual pheno-seq.

In general, the reagent composition in both methods is very similar and consists of optimized components for full-length scRNA-seq based on template switching technology[86]. However, cellular lysis is most likely more effective for manual pheno-seq workflow, which involves strong detergents and subsequent RNA-isolation. In contrast, HT-pheno-seq relies on cellular lysis by freezing/thawing and Triton-X detergent. Although we already increased the detergent concentration 5-fold compared to the single cell protocol, this might have to be further optimized. In addition, the reaction volume is approximately 500-fold lower for HT-pheno-seq, which could in principle improve performance of the RT reaction[77]. At the same

time, the amount of cellular material present during the reaction is increased and could have an inhibitory effect on both RT and cDNA amplification. Whereas manual pheno-seq cDNA libraries are generated by separated lysis, RT and cDNA amplification and enable full-length sequencing, HT-pheno-seq libraries are generated in a one-step PCR and only consist of 3´-ends due to the high multiplexing. A major improvement of HT-pheno-seq could be the usage of unique molecular identifiers (UMIs), random barcodes that enable transcript counting to avoid amplification bias, or the ability to generate full-length libraries. However, the used protocol for the iCELL8 system does not support the generation of UMI or full-length libraries to date. Finally, fixation by DSP seems to have a slight negative influence on either lysis or RT efficiency (Figure 2.2h). Although de-crosslinking is based on Dithiothreitol (DTT), a general component of the reverse transcription buffer, the concentration, temperature and incubation time with DTT might have to be optimized.

3.1.3.3 Pheno-seq imaging

A major advancement of pheno-seq is the direct combination of gene expression profiling and imaging. Whereas library preparation and RNA-sequencing workflows are relatively straightforward and well established, high-throughput microscopy is less well optimized and technically more challenging, especially in this specific setup. A key consideration for pheno-seq is the balance between time and image resolution as increased imaging time that is generally necessary for enhanced image quality can severely affect RNA quality. The current protocol based on confocal laser scanning microscopy takes approximately 30 minutes, which we considered as maximum to not severely affect RNA quality, even in DSP fixed samples. Although we could highly increase spheroid image quality compared to the default iCELL8 imaging system (Figure 2.4), there is still much room for improvement. We envision pheno-seq to become even more powerful with single cell or even subcellular resolution, 3D image acquisition and time-lapse microscopy, as well as integrated staining by IF, live dyes or even RNA-FISH. Whereas heterogeneous spheroid phenotypes of >10-20 cells are distinguishable with the current workflow, significantly enhanced image resolution might resolve ‘spheroid’ phenotypes of very low cell numbers (1-5 cells), which would simultaneously simplify gene expression deconvolution to single cell resolution (Figure 2.12). Increasing pheno-seq image content to that extent requires alternative imaging technology, and light sheet fluorescence microscopy (LSFM)[259–261] most probably represents the only possible solution that meets all requirements. In contrast to point-detection and scanning in confocal microscopy where the whole sample is illuminated throughout imaging, LSFM uses a sheet of light that illuminates the sample only in one thin section perpendicularly to the

detection objective. This 'optical sectioning' principle enables high-speed imaging at high spatial and temporal resolution with minimal photodamage. Very recently, a new generation of LSFM have been introduced, combining lattice light sheet microscopes with adaptive optics, thereby enabling detailed imaging of subcellular events *in vivo*[262]. As proof-of-concept, we were able to image spheroids in the iCELL8 nanowell chip setup with a dual-view inverted selective plane illumination microscope[261] in combination with a fluorinated ethylene propylene (FEP) foil to seal the chip (not shown). Notably, imaging of one row only (72 nanowells) at one time-point can generate data in the range of hundreds of gigabytes. Thus, additional challenges for data processing and storage arise when using this kind of microscopy for pheno-seq. Alternatively, decreasing the number of samples/wells in alternative culture and imaging setups might enable detailed time-lapses of clonal spheroids without producing vast amounts of data.

3.1.3.4 Data integration and analysis

To our knowledge, pheno-seq is the first method that enables direct combined data analysis of both gene expression and contextual image features. However, combined analysis presented in this study does not involve a real integration of imaging and RNA-seq datasets but rather a mapping of image features on pre-defined gene expression clusters and subsequent statistical analysis. Although this represents a significant technical advancement with high biological relevance, this strategy has been mainly chosen due to the low image complexity and future efforts should generally focus on gaining single cell imaging resolution. Besides technological future goals for pheno-seq implementing higher resolution microscopy and content, key advancements will include the direct integration of complex image feature[53] and RNA-seq datasets. Similar to the integration of different (single cell) Omics datasets[263–265], profiling different layers from the same sample will aid to identifying biologically relevant molecular and phenotypic features and their connection. For example, informative image features might improve gene expression clustering and *vice versa*. Moreover, this strategy might help to identify subtype-specific but primarily unobserved phenotypes and associated molecular features.

If more complex pheno-seq imaging datasets are generated and the higher the sample number is, more challenges will arise for efficient and accurate data analysis. Especially for image analysis, precise single cell segmentation and feature extraction represent the biggest hurdles especially for huge datasets (e.g. LSFM-derived). In addition, most available tools have been developed for 2D cultured cells or histological slices and not for 3D objects like spheroids. Recent state-of-the art developments in computer vision tasks have been

primarily dominated by deep-learning algorithms[266] of which convolutional neuronal networks[267] (CNNs) are the most relevant for image analysis. In contrast to 'classical' image analysis methods, CNNs learn to extract relevant profiles/patterns directly from raw image data by multiple non-linear transformations without explicit image segmentation and feature extraction. As CNNs provide several advantages including enhanced speed and improved performance, these approaches appear as most promising to handle complex imaging datasets. For instance, a CNN combined with non-linear dimension reduction has been used to infer cell cycle stages and other single cell phenotypes from high-throughput imaging flow-cytometry, thereby substantially increasing speed and decreasing error rates compared to 'classical' approaches based on feature extraction. Finally, CNNs might even improve RNA-seq and image data integration although this needs to be demonstrated.

3.1.4 Potential extensions of pheno-seq

In addition to above described possible improvements of the current workflow (section 3.1.3), pheno-seq might be also extended to other experimental setups or combined with other imaging and NGS technologies to improve cellular, spatial and temporal resolution or to add other regulatory layers of gene expression.

3.1.4.1 Acquiring single cell resolution at the gene expression level

Pheno-seq enables unbiased detection of heterogeneous gene expression with a higher sensitivity than scRNA-seq, but at cost of lower cellular resolution (Figure 2.2). Although pheno-seq still provides single cell information as the profiled spheroid phenotype and its transcriptome are the direct consequence of its founding cell, the resulting multicellular structure will consist of several distinct cell types or states if the used 3D culture system closely reflects physiological conditions. Thus, the more complex the 3D cell culture system is, and the higher the cell numbers in profiled spheroids are, the more important will it be to acquire 'real' single cell information on the gene expression level.

We could deconvolve CRC pheno-seq data by a maximum likelihood inference approach and identified over 1000 highly relevant genes, including many known stem cell markers which we also identified by scRNA-seq (Supplementary Figure 10), that are likely to exhibit heterogeneous expression at the single cell level (Figure 2.12). However, the deconvolution approach in its current form does not inform about the global relationship of these genes, including gene expression correlation across single cells, which would be needed to specifically define cellular subtypes. Thus, combining pheno-seq with scRNA-seq and integrated data analysis might be a pragmatic strategy to circumvent current limitation in

cellular resolution. For instance, Moncada et al. circumvented the lower cellular resolution of spatial transcriptomics by combining this approach with scRNA-seq from the same tissue piece of pancreatic ductal adenocarcinoma[268]. By combining information regarding inferred cellular subpopulations and gene expression maps, they could deconvolve proportions of cell types that localize to specific regions inside the tissue, including cancer cell subtypes, fibroblasts, normal epithelial cells and different kinds of immune cells. Accordingly, the same strategy could be applied to deconvolve cell type compositions in imaged spheroids.

3.1.4.2 Pheno-seq and time-lapse microscopy

Enhancing image resolution will be one of the key improvements of pheno-seq to ultimately gain single cell or even subcellular resolution (discussed in section 3.1.3.3). A key advantage of *in vitro* cell culture systems is the experimental control over time. This does not only enable sampling at defined time-points and under defined conditions, but also imaging of dynamic processes over time. Therefore, integration of time-lapse microscopy in pheno-seq workflows appears as a key step to understand growth dynamics, changes and appearances of visual cellular phenotypes and, most importantly, underlying gene expression at defined end time-points. For example, pheno-seq gene expression profiling at various time points after single cell seeding and integrated time-lapse microscopy could reveal the link between visual cellular phenotypes, including dynamic behavior, and appearances of subtypes. Alternatively, the end time point for gene expression profiling could be defined by the occurrence of specific cellular events. Moreover, time-resolved microscopy prior to gene expression profiling could be combined with fluorescent reporters for subtype markers that were identified beforehand with scRNA-seq or pheno-seq. Prominent examples of how continuous time-lapse microscopy with fluorescent reporters for subtype markers can change previous assumptions of lineage choice based on single cell or bulk RNA-seq come from studies with hematopoietic stem cells. First, Hoppe et al. could show that different myeloid lineage associated transcription factors are ‘only’ executing and reinforcing lineage choices that are already made, rather than competing against each other[269]. In addition, by using a CNN based on brightfield images, they could later show that lineage choice can be detected up to three generations before known lineage markers are even detectable[270]. As already mentioned before, combining time-lapse microscopy with pheno-seq will not be possible with the current iCELL8 based workflow. However, first trials of experimental setups that are compatible with time-lapse microscopy based on LSFM, including single cells spotted in small drops of Matrigel, show promising results (not shown).

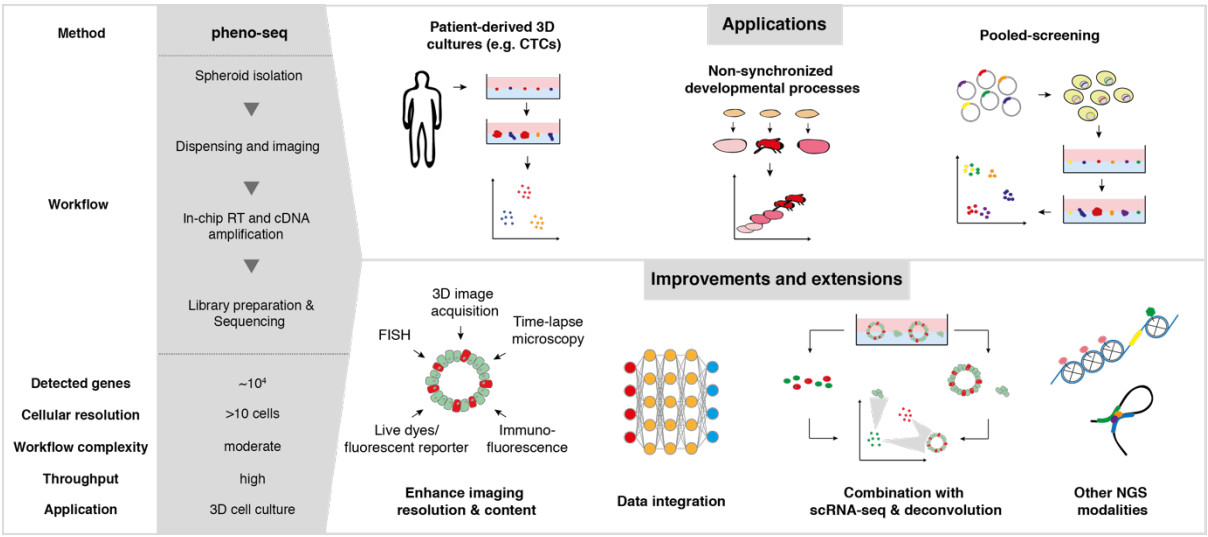


Figure 3.1 I Pheno-seq summary and outlook. Left: Overview and key characteristics of pheno-seq as new method to directly combine unbiased gene expression profiling and microscopy in 3D culture systems (similar to Figure 1.4). Upper right: Possible application strategies for pheno-seq. Lower right: Potential technical and analytical improvements and extensions for pheno-seq.

3.1.4.3 Additional extensions for pheno-seq

In recent years, the ability to perform unbiased gene expression analysis in single cells has led to the development of many other molecular methods that reveal other or additional layers of information in low-input samples or single cells, including single cell (epi)genomic profiling[92,94,95,271,272], combinations of single cell -Omics from the same single cell[96–100,273], as well as lineage tracing[274–277] and functional screening[258,278–280] combined with scRNA-seq. Obviously, many of these methods might be combined with the pheno-seq principle. For instance, pheno-seq can be easily extended to other low-input, next-generation sequencing modalities such as chromatin accessibility prolifiling[95], even with the iCELL8 technology[281]. Furthermore, the combination of lineage tracing with time-lapse microscopy and subsequent gene expression profiling would reveal a direct link between continuous single cell behavior and developmental trajectories, although a ‘real’ single cell readout might be required in this experimental setup. In order to achieve such resolution for direct image and gene expression profiling at the single cell level, technological advances need to either (i) enable the isolation of all cells of one particular spheroid after imaging, or (ii) enable unbiased spatial transcriptomics at single cell resolution, optimally in 3D[135].

3.2 Heterogeneous metabolic signatures are linked to cancer cell differentiation in a 3D model of colorectal cancer

3.2.1 Cellular composition and hierarchical organization in a 3D model of CRC

The intestinal epithelium represents a well characterized hierarchically organized cellular compartment, with self-renewing LGR5$^+$ stem cells at the crypt base giving rise to more specialized and short-lived cellular progeny that differentiate upon moving upwards the crypt. An increasing number of investigations suggests that CRC exhibits a similar hierarchical organization in which LGR5$^+$ cancer stem cells represent the cells of origin fueling the growth of developing adenomas and tumors[164,205,238] (see section 1.3.1.2). However, relatively little is known about the cellular composition and underlying transcriptional programs in CRC and understanding phenotypic characteristics and dependencies of CSCs could reveal potential targets to efficiently treat CRC.

In order to resolve cellular heterogeneity in CRC in more detail, we applied scRNA-seq to 3D spheroid cultures derived from 12 CRC patients. Using the same culture model, Dieter et al. have identified functionally distinct subpopulations with different self-renewal and tumorigenic capacity[204]. These functional differences were independent of underlying subclone composition[226], indicating non-genetic factors that drive tumorigenicity. In addition, we identified transcriptional signatures that correlated with the proliferative potential of clonal CRC spheroids by pheno-seq. These included known stem and differentiation markers, further supporting a differentiation hierarchy in CRC and indicating for multiple functionally different cellular subtypes that relate to known intestinal lineages (see section 2.1).

In line with these observations, we inferred lineage-specific transcriptional signatures by scRNA-seq from multiple patients, including LGR5$^+$ stem cells, potentially niche forming DEFA5$^+$ Paneth cells, a putative transit-amplifying compartment driven by MYC as well as a terminally differentiated subtype (FABP1$^+$/H1F0$^+$). In addition, we identified diverse cell states including metabolic signatures for OXPHOS, hypoxia/glycolysis and fatty acid metabolism as well as signatures associated with cellular proliferation (Figure 2.15). Finally, we identified transcriptional programs linked to immune response (e.g., CCL20, IL-8; CXCL1) and antigen presentation (HLA-E, HLA-F), indicating that immune-modulatory gene expression is maintained in culture and associated with specific cellular subtypes or states[282,283]. As this 3D cell culture model was previously believed to 'only' enrich for CSCs[17,31,204], the degree of heterogeneity and differentiation is surprisingly high and might be much closer to the actual cancer cell composition in primary tumors (or CSC derived organoids) than expected. In line with this, CRC spheroids clearly reflect crypt-like morphological characteristics (Figure 2.17). While further validation is still required, we

derived a potential lineage hierarchy in CRC from the identified transcriptional signatures and previous knowledge of marker genes and lineage relationships of intestinal subtypes[159] (Figure 3.2a): LGR5$^+$ cells represent the CSC pool that gives rise to either DEFA5$^+$ Paneth-like cells or to putative highly cycling transit-amplifying cells (MYC$^+$) that further differentiate into terminally differentiated cells (FABP1$^+$). Moreover, these subtypes are characterized by heterogeneous metabolic states, a phenomenon that has been also observed in the healthy small intestine[192] (see section 3.2.3). However, the accurate determination of lineage relationships most probably requires additional analysis strategies and alternative experimental setups (see section 3.2.2 and 3.2.4). Furthermore, it is important to note that the definitions of and differences between cell 'state' and 'type' are not yet fully understood[38]. Most likely, cell type transitions are characterized by continuous processes that involve multiple cell states and that are influenced by intrinsic and extrinsic factors[129,284].

3.2.2 Challenges and limitations in analyzing scRNA-seq data of cancer patients

The analysis of scRNA-seq data is associated with specific computational and statistical challenges (see section 1.2.2.3). Additional complications arise for scRNA-seq data generated from cancer cells, for example due to patient-specific (epi)genomic alterations. Similar to tSNE and PCA, NNMF analysis did not result in mutually exclusive clusters, as observed for cell types in the healthy small intestine[88], but rather in overlapping signatures and tendencies. This phenomenon has been observed in multiple tumor entities[87,89,243] and could be due to increased transcriptional noise similar to aging cells[285]. Moreover, several studies have shown that differentiated intestinal cell types can de-differentiate into LGR5$^+$ stem cells upon ablation or irradiation in both tumor and healthy tissue[166,255,286]. This plasticity could be deregulated in CRC due to lowered regulatory barriers between lineages, leading to ongoing interconversion between subtypes and thus to a higher occupancy of transition states that might blur cellular clustering. The co-expression of PROX1 with LGR5 provides another indication of mis-regulated lineage plasticity, as PROX1 is normally expressed in cells of the enteroendocrine lineage that revert into LGR5$^+$ stem cells upon injury in normal intestinal tissue[232].

Although the total number of cells in this study is relatively high (nearly 5000), the number of cells per patient is relatively low (approximately 400) which limited the detailed analysis of individual patients. Thus, profiling of greater numbers of tumor cells will be important for further analysis, including improved clustering and the identification of lineage relationships between cellular subtypes[287,288].

Furthermore, in our approach we combined data from multiple patients in order to identify shared expression programs across patients. Although mean-centering represents a straightforward and widely used method to circumvent inter-patient variability, it might be limited in integrating unique mutational signatures. Several alternative methods have been developed for technical or biological batch effect correction[115,289,290], but are more tailored to combining datasets from different technologies or species. Thus, more sophisticated strategies will be needed to correct for inter-patient variability, potentially by using approaches that 'learn' cellular subtypes from healthy cells of the same tissue[291]. Adding information from bulk or single cell DNA sequencing will also be crucial to understanding the link between genomic alterations, subclone composition and single cell transcriptomes. In addition, other single cell sequencing modalities for epigenetic[95,271] or multimodal profiling[96,98,292] will further extend our understanding of regulatory gene expression networks and their perturbations during tumorigenic progression.

3.2.3 Metabolic heterogeneity in CRC

Besides the lineage-related cancer cell subtypes detected in our study, heterogeneous metabolic gene expression programs represented a major source of variability across CRC patients. The anti-correlated OXPHOS and glycolysis/hypoxia signatures showed the most pronounced patterns and were mainly associated with the highly cycling MYC^+ compartment and $FABP1^+/H1F0^+$ terminally differentiated cells, respectively (Figure 2.14, Figure 2.15 and Figure 2.16). More detailed analysis of other lineage-specific transcriptional signatures revealed similar metabolic trends, showing that stem-like cells preferentially express OXPHOS genes, whereas Paneth-like cells showed higher expression of genes involved in glycolysis/hypoxia (Figure 2.16). These differential metabolic signatures have also been observed in the intestinal stem cell niche, where Paneth cells favor glycolysis and fuel OXPHOS in stem cells by providing lactate[192].

In general, high rates of OXPHOS can lead to enhanced generation of reactive oxygen species (ROS) that contribute to cellular oxidative stress. However, it has been shown that intestinal stem cells do not exhibit enhanced cytoplasmic ROS compared to Paneth cells despite higher rates of OXPHOS[192]. Although no compensatory mechanism was described, the authors could show that enhanced ROS induce differentiation in stem cells by activating intracellular signaling involving p38. In the scRNA-seq data presented in this work, the meta-signatures for OXPHOS and stem cells both contain genes involved in the protection against oxidative stress, which might represent a compensatory mechanism for high intracellular ROS levels in highly cycling and stem cells of CRC. Whereas the OXPHOS signature is

associated with expression of the thiol-specific peroxidases PRDX3 and PRDX4 (Supplementary Figure 10), the stem-like signature is associated with enhanced expression of OXR1 and PON2, genes essential for protection against oxidative stress[250,251], Glutamate Ammonia Ligase (GLUL) that catalyzes the synthesis of glutamine (a major source for OXPHOS[252]) as well as MAP2K6, an essential component of the p38 signaling cascade[253].

Overall, our scRNA-seq data thus reveals a tight link between cellular differentiation and metabolism in CRC and indicates the preservation of similar mechanisms as observed in healthy intestinal cells. Notably, this link is not restricted to stem and Paneth-like cells but also extends to other subtypes. For example, the Tdiff FABP1^{+} subtype exhibits the highest expression of genes involved in glycolysis (Figure 2.16) and therefore might also provide lactate to stem-like cells or even to highly cycling TA (MYC^{+}) cells. In addition, the glycolysis/hypoxia signature contains the gene vascular endothelial growth factor A (VEGFA), a secreted factor known to be involved in signaling that sustains survival and proliferation in CSCs[293]. Thus, glycolytic/hypoxic signaling of differentiated CRC cells might support CSCs beyond providing metabolic products (previous section summarized in Figure 3.2).

Metabolic heterogeneity and dependencies of CSCs have been also described for other tumor entities, including pancreatic cancer[294], breast cancer[198], glioma[295] and leukemia[296], indicating a widespread phenomenon. Metabolic preferences of CSCs depend on the tumor entity and might be characterized by a high degree of plasticity in order to adapt to changing environments[155,295]. In addition, analysis of intrinsically driven metabolic identities in primary tumors might be confounded by environmental influences, such as hypoxic regions. *In vitro* cell culture systems provide an alternative strategy to analyze metabolic heterogeneity as they enable control over environmental conditions, but most studies so far have used cells that were cultured under high glucose and oxygen conditions, thus favoring glycolysis. Notably, the 3D culture system utilized in this study exhibits a high degree of metabolic heterogeneity despite culture conditions that favor glycolysis, indicating that intrinsic metabolic preferences of inferred intestinal lineages are hardwired in CRC.

Optimally, single cell gene expression profiles should be complemented by single cell metabolic profiles in order to assess whether detected gene expression heterogeneity results in functionally different metabolic profiles. However, most quantitative methods for single cell metabolomics are relatively new and immature[201]. Alternatively, subpopulations identified by scRNA-seq could be isolated by FACS to acquire enough material for accurate metabolic profiling.

3.2.4 Niche dependencies, metabolic compartments and the future of *in situ* analysis

The intestinal stem cell niche is characterized by specific cellular interdependencies, in which stem cells compete with each other for cellular surface of adjacent Paneth cells which sustain the niche by providing growth factors[4,229] and metabolic products[192]. This interdependency seems to be maintained in developing adenomas where LGR5$^+$ stem cells are frequently localized adjacent to Paneth cells[164]. In this work, we could detect both stem (LGR5$^+$) and Paneth-like (DEFA5$^+$) transcriptomic signatures in 3D cultures of advanced CRC (Figure 2.15 and Supplementary Figure 10), suggesting similar interactions.

In situ analysis by microscopy did not only validate results obtained by scRNA-seq in terms of identified subpopulations (Figure 2.18 and Supplementary Figure 14), but also provided further insight into the spatial distribution of subtype and states. For example, we could frequently detect LGR5$^+$ and DEFA5$^+$ in close proximity to each other in all three patients (Figure 2.18 and Supplementary Figure 14), thus indicating for niche promoting effects of the Paneth-like subpopulation similar to the healthy intestine. This notion is supported by the fact that the Paneth-like expression signature contained the secreted and WNT-pathway related factors Midkine (MDK)[297] and FRZB[298], which might modulate niche characteristics of Paneth-like cells in CRC. Similarly, niche-promoting subpopulations that express WNT and NOTCH components have been identified in different lung cancer entities[170,171], suggesting that niche-specific cellular interactions are a central feature of tumor progression across different cancer entities. However, the strict spatial distribution of intestinal cell types in the healthy crypt is strongly perturbed in CRC spheroids (Figure 2.17), which is most probably associated with partial loss of niche dependencies driven by perturbations of WNT pathway signaling[155].

Additionally, we could detect strong differences in numbers of mitochondria between single cells by using a mitochondrial live-dye (Figure 2.17). As the abundance of mitochondria is correlated with increased respiratory activity in intestinal organoids, we reasoned that cells with high numbers of mitochondria represent the same cells that exhibit high expression of the OXPHOS expression signature identified by scRNA-seq (defined as OXPHOS^{high}). Most intriguingly, OXPHOS^{high} regions are specifically located at the outer layer of crypt-like regions in all three patients (Figure 2.17). This observation has striking similarities to healthy intestinal organoids (Figure 3.2b), where crypt formation ('budding') and associated differentiation are driven by OXPHOS and ROS signaling[192]. As OXPHOS and cell cycle signatures exhibit a high overlap in scRNA-seq data, OXPHOS^{high} compartmentalization could be associated with continuous crypt budding and proliferation of high-cycling MYC$^+$ cells at the outer regions of CRC spheroids, which might enable tumor growth *in vivo*.

Notably, histological classification and quantification of tumor budding regions serves as prognostic marker for CRC[299].

Furthermore, anti-correlated metabolic tendencies of stem and Paneth-like cells inferred by scRNAseq could be validated *in situ*, showing that LGR5^{+} cells exhibit a more pronounced overlap with OXPHOS^{high} cells at putative budding regions (Figure 2.19, Figure 2.21 and Supplementary Figure 15). It is important to note that observed metabolic identities in OXPHOS and glycolysis are not mutually exclusive for stem and Paneth-like cells in both scRNA-seq and *in situ* data, but rather represent strong tendencies. This is most likely associated with the spatial location of subtypes in spheroids, leading to variable metabolic programs in both subtypes. For example, we also observed LGR5^{+} cells in OXPHOS^{low} regions in the inner part of spheroids (Supplementary Figure 16) which might represent quiescent CSCs that have been described previously[165].

Nevertheless, these results suggest a central role of stem-like cells and their associated metabolic state in driving tumor cell growth at peripheral tumor regions which might be supported by Paneth and Tdiff cells. Due to the similarity to the maturation of healthy intestinal organoids from spherical embryonic-like spheroids[192], it is likely that CRC cells hijack developmental programs to sustain tumor growth as observed in other tumor entities[89,291]. However, it remains unclear which mechanisms drive crypt budding in CRC spheroids. Although this needs to be further validated, cellular interactions and mechanical forces might influence these processes.

In order to analyze cellular interactions and dynamic processes in more detail, several improvements and alternative experimental setups will be necessary in the future. First, the number of profiled genes by RNA-FISH is highly restricted by the current approach and limits the number of cellular subtypes and states that can be mapped at the same time. Thus, utilizing methods for high multiplexing of RNA-FISH probes[58,130,300] will enable a much more detailed view on the spatial organization and associated genetic programs in CRC spheroids. Second, RNA-FISH staining of intact spheroids and 3D image acquisition will avoid confounding effects of histological preparation and enable a deeper understanding of spheroid morphology, cellular architecture and spatial localization of subtypes. Finally, both scRNA-seq and RNA-FISH only provide a snap-shot of gene expression and does not inform about dynamic processes the are inherent to cellular behavior. Therefore, multi-color 3D time-lapse microscopy with metabolic live-dyes and fluorescent-reporters for lineage marker genes will answer remaining questions about dynamic metabolic states, lineage plasticity and mechanical forces as well their contribution to crypt budding and proliferative behavior in CRC spheroids.

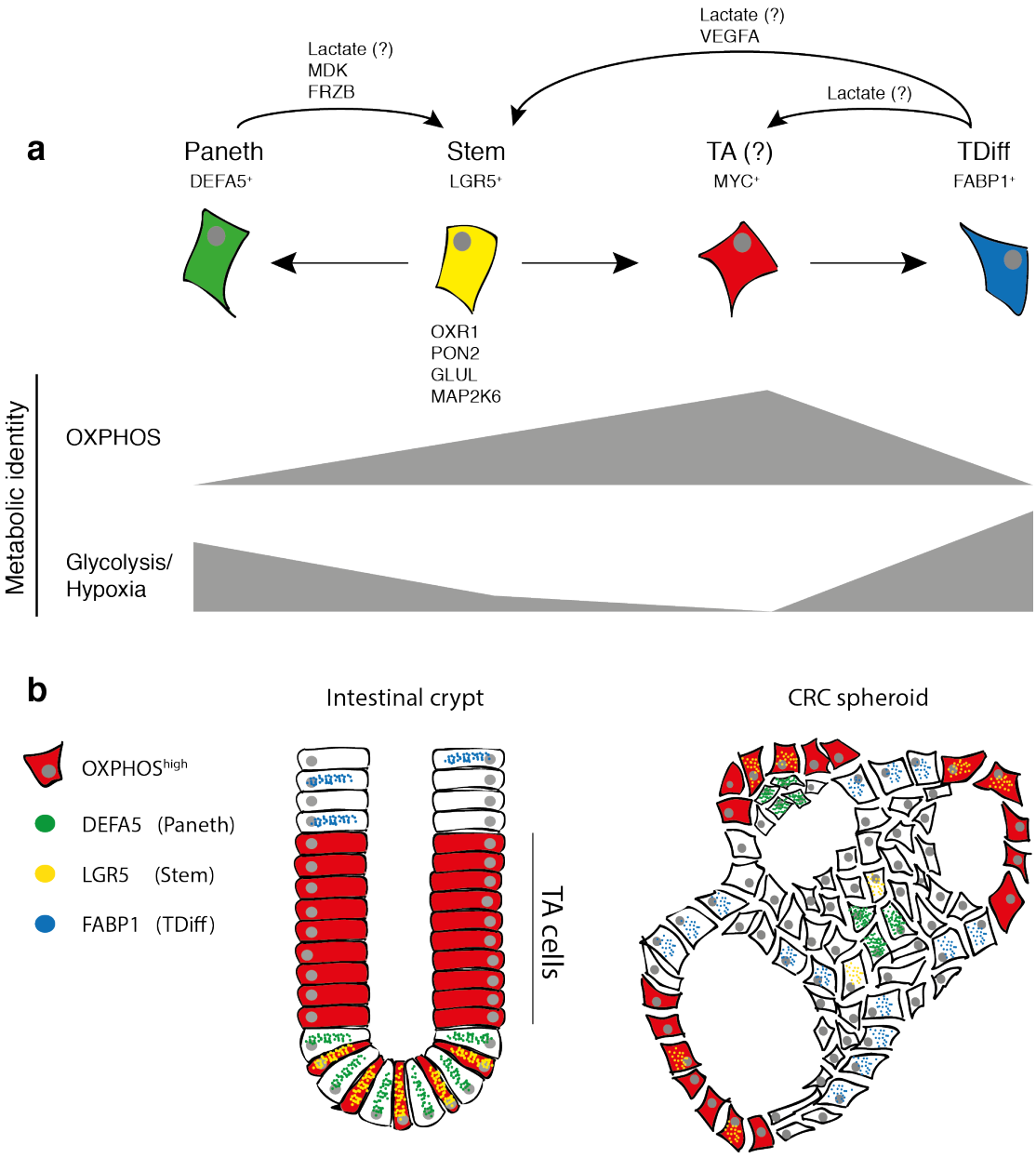


Figure 3.2 | Schematic summary of the main findings regarding cancer cell heterogeneity in CRC. **(a)** Potential lineage relationships and potential interactions between identified CRC subtypes as well as associated metabolic preferences. **(b)** Schematic intestinal crypt (left) and CRC spheroid (right). Representative spatial distribution of lineage subtypes and mitochondria as inferred by quantitative *in situ* analysis using RNA-FISH and Mitotracker live-dye in CRC spheroids.

3.2.5 Functional analysis of tumor cell heterogeneity and metabolic states

The results presented in this work demonstrate a surprisingly high degree of heterogeneity in the analyzed patient-derived 3D model of CRC, including interconnections between intestinal lineages, metabolic states and spatial locations. However, the results are mainly descriptive and thus do not demonstrate functional implications of identified subtype characteristics. Therefore, future work will need to address the question whether identified subtypes, overlapping cell states and associated genes can be linked to functional differences, for example in proliferative and tumorigenic potential. Pheno-seq results provide first indications regarding the proliferative behavior of identified CRC subtypes, as genes associated with the putative LGR5^{+} CSC subpopulation identified by scRNA-seq show a clear overlap with genes that are associated with high proliferative capacity (e.g., LGR5, SMOC2, PROX1, APP, ITM2B). Moreover, pheno-seq indicates for a heterogeneous growth phenotype of DEFA5^{+} cells that we also identified by scRNA-seq (see section 2.1.4), which might indicate for cancer cell plasticity and de-differentiation of Paneth into stem-like CRC cells, potentially initiated by autocrine signaling if a certain number or density of DEFA5^{+} cells is reached.

The most striking questions regarding functional implications of identified cancer cell heterogeneity refer to lineage-specific metabolic preferences as they may provide promising targets to eradicate colorectal CSCs. The identified metabolic patterns are closely resembling those occurring in the healthy intestine and future experiments will need to address the question whether similar mechanisms and dependencies drive CSC maintenance and cancer cell differentiation. For example, lactate might also be a source for OXPHOS in stem-like cells that is provided by glycolytic Paneth and Tdiff cells, and the expression of MAP2K6 in the stem expression signature indicates for p38-dependent cancer cell differentiation in CRC. Moreover, stem-like cells seem to protect themselves against oxidative stress by expression of OXR1 and PON2, indicating that these genes are required for CSC maintenance. Hence, perturbation assays by targeting metabolic pathways, p38 signaling and genes protecting against ROS will provide more information on functional metabolic states in CRC.

Optimally, proliferative capacity and tumorigenic potential of identified cellular subtypes and states should be analyzed independently for each subpopulation, which involves sorting of identified subtypes based on surface marker heterogeneity or by engineering of fluorescent reporter lines for specific marker genes. In addition, live-dyes can used to selectively stain cells with distinct metabolic states. First results of sorted and replated Mitotracker^{high} and Mitotracker^{low} cells revealed strongly enhanced spheroid forming capacity in the Mitotracker^{high} compartment (not shown), thereby indicating for functional implications of

heterogeneous metabolic states in CRC. Ultimately, results obtained from *in vitro* assays will need to be validated in serial xenograft transplantations or lineage tracing experiments *in vivo*.

3.2.6 Reliability and limitations of 3D cell culture models to reflect primary tumors

During the last years, 3D cell culture systems have evolved as promising tools to model primary tumor cell behavior and drug responses[17,32]. For example, the spheroid cultures utilized in this study have been derived from CRC patients and maintain many features or primary tumors, including intestinal morphologies (Figure 2.17), genomic alterations and subclone architecture[226], inter-patient variability in gene expression (Figure 2.13), hierarchical organization[204] and lineage-specific gene expression (Figure 2.15) as well as gene expression that is associated with immune-modulatory function (Supplementary Figure 10).

However, flexibility of *in vitro* systems comes at cost of reduced representation of the tumor microenvironment. Initial studies have recently started to model and improve microenvironmental influences in 3D culture systems, including ECM[5], vascularization[301], fibroblasts[302] or peripheral blood lymphocytes[303], but these models are still limited in comprehensively reflecting all environmental influences[304]. In addition, 3D culture systems underlie several artificial influences, including the composition of applied media as well as cellular dissociation and passaging, which can lead to substantial differences in cultivation efficiencies and potentially in the loss of cellular subclones. Thus, future studies will need to focus on the analysis of reliability of 3D cell culture systems compared to primary samples, optimally by single cell approaches[207].

3.3 Conclusion

In vitro cell culture systems are inevitably necessary models for human disease and the more physiologic they are, the more predictive they will be for personalized treatments. Novel 3D cell culture systems have recently found their way into translational research and clinical settings, where they might become standard tools if combined with state-of-the-art molecular profiling and computational analysis (https://lifetime-fetflagship.eu). However, in order to use them as patient-specific 'avatars', these *in vitro* cultures should optimally be analyzed in the same depth as primary samples, which holds especially true for the analysis of intratumor heterogeneity. Whereas single cell approaches based on NGS have already started to transform our understanding of tumor cell heterogeneity in primary samples, these technological advances have not been broadly transferred to *in vitro* models, although they provide much more experimental flexibility.

With this work, we provide two complementary strategies to demonstrate the power of single cell analysis in 3D cell culture systems by combining quantitative imaging and NGS. First, pheno-seq directly links functional single cell growth phenotypes and morphological features with transcriptional programs, which represents a complementary strategy to spatial transcriptomics in primary tumor samples. Second, we used an indirect approach by applying scRNA-seq to 3D cultures of multiple patients, integrating acquired data to identify shared expression programs and extending results by mapping identified markers to spatial locations by RNA-FISH. As both methods complement each other, we envision that the combination of both approaches across multiple patients evolves as standard strategy to analyze tumor cell heterogeneity in 3D cell culture systems. For example, pheno-seq could be used to further define budding phenotypes of CRC spheroids and associated dynamic transcriptional programs. Meanwhile, especially the applied imaging-based methods will require significant improvements, including higher imaging resolution and content. In addition, these two approaches represent initial strategies to understand the underlying cell culture systems in detail and should provide the basis for further functional experiments, for example *in vivo* lineage tracing with identified markers. Finally, 3D cell cultures systems are still limited in reflecting all components of primary tumors. Thus, the optimal approach to tackle intratumor heterogeneity will involve the integration of results obtained from both primary tumor samples and *in vitro* cultured cells in order to assess the predictive power of the used cell culture system.

4 References

1. Taylor, M. W. in *A History of Cell Culture* **3,** 41–52 (2014).
2. Landry, J. J. M. *et al.* The Genomic and Transcriptomic Landscape of a HeLa Cell Line. *G3 Betheda* **3,** 1213–1224 (2013).
3. Streuli, C. H., Bailey, N. & Bissell, M. J. Control of mammary epithelial differentiation - basement-membrane induces tissue-specific gene-expression in the absence of cell cell-interaction and morphological polarity. *J. Cell Biol.* **115,** 1383–1395 (1991).
4. Sato, T. *et al.* Paneth cells constitute the niche for Lgr5 stem cells in intestinal crypts. *Nature* **469,** 415–418 (2011).
5. Gjorevski, N. *et al.* Designer matrices for intestinal stem cell and organoid culture. *Nature* **539,** 560–564 (2016).
6. Lane, S. W., Williams, D. A. & Watt, F. M. Modulating the stem cell niche for tissue regeneration. *Nat. Biotechnol.* **32,** 795–803 (2014).
7. Wang, F. *et al.* Reciprocal interactions between 1-integrin and epidermal growth factor receptor in three-dimensional basement membrane breast cultures: A different perspective in epithelial biology. *Proc. Natl. Acad. Sci.* **95,** 14821–14826 (1998).
8. Jabs, J. *et al.* Screening drug effects in patient-derived cancer cells links organoid responses to genome alterations. *Mol. Syst. Biol.* **13,** 955 (2017).
9. Lasfargues, E. Y. Cultivation and behavior in vitro of the normal mammary epithelium of the adult mouse. II. Observations on the secretory activity. *Exp. Cell Res.* **13,** 553–562 (1957).
10. Gahmberg, C. G. & Hakomori, S. I. Altered growth behavior of malignant cells associated with changes in externally labeled glycoprotein and glycolipid. *Proc Natl Acad Sci U S A* **70,** 3329–3333 (1973).
11. Timpl, R. *et al.* Laminin - A glycoprotein from basement membranes. *J. Biol. Chem.* **254,** 9933–9937 (1979).
12. Orkin, R. W. *et al.* A murine tumor producing a matrix of basement membrane. *J. Exp. Med.* **145,** 204–20 (1977).
13. Lee, E. Y., Lee, W. H., Kaetzel, C. S., Parry, G. & Bissell, M. J. Interaction of mouse mammary epithelial cells with collagen substrata: regulation of casein gene expression and secretion. *Proc.Natl.Acad.Sci.U.S.A* **82,** 1419–1423 (1985).
14. Li, M. L. *et al.* Influence of a reconstituted basement membrane and its components on casein gene expression and secretion in mouse mammary epithelial cells. *Proc. Natl. Acad. Sci.* **84,** 136–140 (1987).

15. Barcellos-Hoff, M. H., Aggeler, J., Ram, T. G. & Bissell, M. J. Functional differentiation and alveolar morphogenesis of primary mammary cultures on reconstituted basement membrane. *Development* **105,** 223–235 (1989).
16. Petersen, O. W., Ronnov-Jessen, L., Howlett, A. R. & Bissell, M. J. Interaction with basement membrane serves to rapidly distinguish growth and differentiation pattern of normal and malignant human breast epithelial cells. *Proc. Natl. Acad. Sci.* **89,** 9064–9068 (1992).
17. Weiswald, L. B., Bellet, D. & Dangles-Marie, V. Spherical Cancer Models in Tumor Biology. *Neoplasia (United States)* **17,** 1–15 (2015).
18. Takahashi K, Tanabe K, Ohnuki M, Narita M, Ichisaka T, Tomoda K, Y. S. Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell.* **131,** 861–72 (2007).
19. Ma, L., Renner, M. & Ca, M. Cerebral Organoids Model Human Brain Development and Microcephaly. *Mov. Disord.* **29,** 25740 (2014).
20. Barker, N. *et al.* Identification of stem cells in small intestine and colon by marker gene Lgr5. *Nature* **449,** 1003–1008 (2007).
21. Sato, T. *et al.* Single Lgr5 stem cells build crypt-villus structures in vitro without a mesenchymal niche. *Nature* **459,** 262–265 (2009).
22. Grün, D. *et al.* Single-cell messenger RNA sequencing reveals rare intestinal cell types. *Nature* **525,** 251–255 (2015).
23. Sato, T. *et al.* Long-term expansion of epithelial organoids from human colon, adenoma, adenocarcinoma, and Barrett's epithelium. *Gastroenterology* **141,** 1762–1772 (2011).
24. Barker, N. *et al.* Lgr5(+ve) stem cells drive self-renewal in the stomach and build long-lived gastric units in vitro. *Cell Stem Cell* **6,** 25–36 (2010).
25. Huch, M. *et al.* Unlimited in vitro expansion of adult bi-potent pancreas progenitors through the Lgr5/R-spondin axis. *EMBO J.* **32,** 2708–2721 (2013).
26. Huch, M. *et al.* In vitro expansion of single Lgr5 + liver stem cells induced by Wnt-driven regeneration. *Nature* **494,** 247–250 (2013).
27. Torre, L. A. *et al.* Global cancer statistics, 2012. *CA. Cancer J. Clin.* **65,** 87–108 (2015).
28. Bedard, P. L., Hansen, A. R., Ratain, M. J. & Siu, L. L. Tumour heterogeneity in the clinic. *Nature* **501,** 355–364 (2013).
29. Kamb, A. What's wrong with our cancer models? *Nat. Rev. Drug Discov.* **4,** 161–165 (2005).

30. Singh, S. K. & Clarke, I. D. Identification of a cancer stem cell in human brain tumors. *Cancer Res* **63,** 5821–8 (2003).
31. Ricci-Vitiani, L. *et al.* Identification and expansion of human colon-cancer-initiating cells. *Nature* **445,** 111–115 (2007).
32. Drost, J. & Clevers, H. Organoids in cancer research. *Nat. Rev. Cancer* 1–12 (2018). doi:10.1038/s41568-018-0007-6
33. Fujii, M. *et al.* A Colorectal Tumor Organoid Library Demonstrates Progressive Loss of Niche Factor Requirements during Tumorigenesis. *Cell Stem Cell* **18,** 827–838 (2016).
34. Van De Wetering, M. *et al.* Prospective derivation of a living organoid biobank of colorectal cancer patients. *Cell* **161,** 933–945 (2015).
35. Sachs, N. *et al.* A Living Biobank of Breast Cancer Organoids Captures Disease Heterogeneity. *Cell* **172,** 373–386.e10 (2018).
36. Drost, J. *et al.* Sequential cancer mutations in cultured human intestinal stem cells. *Nature* **521,** 43–47 (2015).
37. Shimokawa, M. *et al.* Visualization and targeting of LGR5 + human colon cancer stem cells. *Nature* **545,** 187–192 (2017).
38. Regev, A. *et al.* The human cell atlas. *Elife* **6,** 1–30 (2017).
39. Harris, H. *The birth of the cell, Yale Univ. Press, London*. (Yale University Press, 1999).
40. Weigert, R., Porat-Shliom, N. & Amornphimoltham, P. Imaging cell biology in live animals: Ready for prime time. *J. Cell Biol.* **201,** 969–979 (2013).
41. Azaripour, A. *et al.* A survey of clearing techniques for 3D imaging of tissues with special reference to connective tissue. *Prog. Histochem. Cytochem.* **51,** 9–23 (2016).
42. Arthur, G. Albert Coons: harnessing the power of the antibody. *Lancet Respir. Med.* **4,** 181–182 (2016).
43. Köhler, G. & Milstein, C. Continuous cultures of fused cells secreting antibody of predefined specificity. *Nature* **256,** 495–497 (1975).
44. Langer-safer, P. R., Levine, M. & Ward, D. C. Immunological method for mapping genes on Drosophila polytene chromosomes. *Pnas* **79,** 4381–4385 (1982).
45. Heimstädt, O. Das Fluoreszenzmikroskop. *Z. Wiss. Mikrosk.* **28,** 330–337 (1911).
46. Coons, A. H., Creech, H. J. & Jones, R. N. Immunological Properties of an Antibody Containing a Fluorescent Group. *Exp. Biol. Med.* **47,** 200–202 (1941).
47. Specht, E. A., Braselmann, E. & Palmer, A. E. A Critical and Comparative Review of Fluorescent Tools for Live-Cell Imaging. *Annu. Rev. Physiol.* **79,** 93–117 (2017).

48. Chalfie, M., Tu, Y., Euskirchen, G., Ward, W. W. & Prasher, D. C. Green fluorescent protein as a marker for gene expression. *Science (80-.).* **263,** 802–805 (1994).
49. Ploem, J. S. The use of a vertical illuminator with interchangeable dichroic mirrors for fluorescence microscopy with incident light. *Z. Wiss. Mikrosk.* **68,** 129–142 (1967).
50. Wilke, V. Optical scanning microscopy—The laser scan microscope. *Scanning* **7,** 88–96 (1985).
51. Denk *et al.* Two-photon laser scanning fluorescence microscopy. *Science (80-.).* **248,** 73–76 (1990).
52. Huisken, J. & Stainier, D. Y. R. Selective plane illumination microscopy techniques in developmental biology. *Development* **136,** 1963–1975 (2009).
53. Caicedo, J. C. *et al.* Data-analysis strategies for image-based cell profiling. *Nat. Methods* **14,** 849–863 (2017).
54. Boutros, M., Heigwer, F. & Laufer, C. Microscopy-Based High-Content Screening. *Cell* **163,** 1314–1325 (2015).
55. Mannack, L. V. J. C., Eising, S. & Rentmeister, A. Current techniques for visualizing RNA in cells. *F1000Research* **5,** 775 (2016).
56. Wang, F. *et al.* RNAscope: A novel in situ RNA analysis platform for formalin-fixed, paraffin-embedded tissues. *J. Mol. Diagnostics* **14,** 22–29 (2012).
57. Lyubimova, A. *et al.* Single-molecule mRNA detection and counting in mammalian tissue. *Nat. Protoc.* **8,** 1743–1758 (2013).
58. Chen, K. H., Boettiger, A. N., Moffitt, J. R., Wang, S. & Zhuang, X. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science (80-.).* **348,** aaa6090 (2015).
59. Giesen, C. *et al.* Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry. *Nat. Methods* **11,** 417–422 (2014).
60. Marinov, G. K. *et al.* From single-cell to cell-pool transcriptomes: Stochasticity in gene expression and RNA splicing. *Genome Res.* **24,** 496–510 (2014).
61. Ho, B., Baryshnikova, A. & Brown, G. W. Unification of Protein Abundance Datasets Yields a Quantitative Saccharomyces cerevisiae Proteome. *Cell Syst.* **6,** 192–205.e3 (2018).
62. Wang, Y. & Navin, N. E. Advances and Applications of Single-Cell Sequencing Technologies. *Mol. Cell* **58,** 598–609 (2015).
63. Cheung, M. C. *et al.* Intracellular protein and nucleic acid measured in eight cell types using deep-ultraviolet mass mapping. *Cytom. Part A* **83 A,** 540–551 (2013).
64. Eberwine, J. *et al.* Analysis of gene expression in single live neurons. *Proc. Natl. Acad. Sci.* **89,** 3010–3014 (1992).

65. Bengtsson, M., Hemberg, M., Rorsman, P. & Ståhlberg, A. Quantification of mRNA in single cells and modelling of RT-qPCR induced noise. *BMC Mol. Biol.* **9,** 63 (2008).
66. Hoheisel, J. D. Microarray technology: Beyond transcript profiling and genotype analysis. *Nat. Rev. Genet.* **7,** 200–210 (2006).
67. Metzker, M. L. Sequencing technologies the next generation. *Nat. Rev. Genet.* **11,** 31–46 (2010).
68. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: A revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10,** 57–63 (2009).
69. Trapnell, C. Defining cell types and states with single-cell genomics. *Genome Res.* **25,** 1491–1498 (2015).
70. Tang, F. *et al.* mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods* **6,** 377–382 (2009).
71. Kolodziejczyk, A. A., Kim, J. K., Svensson, V., Marioni, J. C. & Teichmann, S. A. The Technology and Biology of Single-Cell RNA Sequencing. *Mol. Cell* **58,** 610–620 (2015).
72. Haque, A., Engel, J., Teichmann, S. A. & Lönnberg, T. A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome Med.* **9,** 75 (2017).
73. Fulwyler, M. J. Electronic separation of biological cells by volume. *Science (80-.).* **150,** 910–911 (1965).
74. Dittrich, W. & Göhde, W. Flow-through chamber for photometers to measure and count particles in a dispersion medium. 1–6 (1973). at <https://www.google.com/patents/US3761187>
75. Chao, M. P., Seita, J. & Weissman, I. L. Establishment of a normal hematopoietic and leukemia stem cell hierarchy. *Cold Spring Harb. Symp. Quant. Biol.* **73,** 439–449 (2008).
76. Jaitin, D. A. *et al.* Massively parallel single cell RNA-Seq for marker-free decomposition of tissues into cell types. *Science (80-.).* **343,** 776–779 (2014).
77. Wu, A. R. *et al.* Quantitative assessment of single-cell RNA-sequencing methods. *Nat. Methods* **11,** 41–46 (2014).
78. Svensson, V. *et al.* Power analysis of single-cell rnA-sequencing experiments. *Nat. Methods* **14,** 381–387 (2017).
79. Pollen, A. A. *et al.* Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. *Nat. Biotechnol.* **32,** 1053–1058 (2014).

80. Klein, A. M. *et al.* Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161,** 1187–1201 (2015).
81. Macosko, E. Z. *et al.* Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161,** 1202–1214 (2015).
82. Gierahn, T. M. *et al.* Seq-Well: Portable, low-cost rna sequencing of single cells at high throughput. *Nat. Methods* **14,** 395–398 (2017).
83. Goldstein, L. D. *et al.* Massively parallel nanowell-based single-cell gene expression profiling. *BMC Genomics* **18,** 1–10 (2017).
84. Goldstein, L. D. *et al.* Massively parallel nanowell-based single-cell gene expression profiling. *BMC Genomics* **18,** 519 (2017).
85. Ramsköld, D. *et al.* Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nat. Biotechnol.* **30,** 777–782 (2012).
86. Picelli, S. *et al.* Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat. Methods* **10,** 1096–1100 (2013).
87. Tirosh, I. *et al.* Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science (80-.).* **352,** 189–196 (2016).
88. Haber, A. L. *et al.* A single-cell survey of the small intestinal epithelium. *Nature* **551,** 333–339 (2017).
89. Tirosh, I. *et al.* Single-cell RNA-seq supports a developmental hierarchy in human oligodendroglioma. *Nature* **539,** 309–313 (2016).
90. Gierahn, T. M. *et al.* Seq-Well: Portable, low-cost rna sequencing of single cells at high throughput. *Nat. Methods* **14,** 395–398 (2017).
91. Rosenberg, A. B. *et al.* Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science (80-.).* **360,** 176–182 (2018).
92. Wang, Y. *et al.* Clonal evolution in breast cancer revealed by single nucleus genome sequencing. *Nature* **512,** 155–160 (2014).
93. Clark, S. J. *et al.* Genome-wide base-resolution mapping of DNA methylation in single cells using single-cell bisulfite sequencing (scBS-seq). *Nat. Protoc.* **12,** 534–547 (2017).
94. Mooijman, D., Dey, S. S., Boisset, J. C., Crosetto, N. & Van Oudenaarden, A. Single-cell 5hmC sequencing reveals chromosome-wide cell-to-cell variability and enables lineage reconstruction. *Nat. Biotechnol.* **34,** 852–856 (2016).
95. Buenrostro, J. D. *et al.* Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523,** 486–490 (2015).
96. Angermueller, C. *et al.* Parallel single-cell sequencing links transcriptional and

epigenetic heterogeneity. *Nat. Methods* **13,** 229–232 (2016).

97. Peterson, V. M. *et al.* Multiplexed quantification of proteins and transcripts in single cells. *Nat. Biotechnol.* **35,** 936–939 (2017).
98. Macaulay, I. C. *et al.* G&T-seq: Parallel sequencing of single-cell genomes and transcriptomes. *Nat. Methods* **12,** 519–522 (2015).
99. Stoeckius, M. *et al.* Simultaneous epitope and transcriptome measurement in single cells. *Nat. Methods* **14,** 865–868 (2017).
100. Clark, S. J. *et al.* Joint Profiling Of Chromatin Accessibility, DNA Methylation And Transcription In Single Cells. *bioRxiv* 138685 (2017). doi:10.1101/138685
101. Stegle, O., Teichmann, S. A. & Marioni, J. C. Computational and analytical challenges in single-cell transcriptomics. *Nat. Rev. Genet.* **16,** 133–145 (2015).
102. Van Den Brink, S. C. *et al.* Single-cell sequencing reveals dissociation-induced gene expression in tissue subpopulations. *Nat. Methods* **14,** 935–936 (2017).
103. Grün, D., Kester, L. & Van Oudenaarden, A. Validation of noise models for single-cell transcriptomics. *Nat. Methods* **11,** 637–640 (2014).
104. Buettner, F. *et al.* Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nat. Biotechnol.* **33,** 155–160 (2015).
105. McCarthy, D. J., Campbell, K. R., Lun, A. T. L. & Wills, Q. F. Scater: Pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. *Bioinformatics* **33,** 1179–1186 (2017).
106. Ilicic, T. *et al.* Classification of low quality cells from single-cell RNA-seq data. *Genome Biol.* **17,** 29 (2016).
107. Vallejos, C. A., Risso, D., Scialdone, A., Dudoit, S. & Marioni, J. C. Normalizing single-cell RNA sequencing data: Challenges and opportunities. *Nat. Methods* **14,** 565–571 (2017).
108. Brennecke, P. *et al.* Accounting for technical noise in single-cell RNA-seq experiments. *Nat. Methods* **10,** 1093–1098 (2013).
109. Kharchenko, P. V., Silberstein, L. & Scadden, D. T. Bayesian approach to single-cell differential expression analysis. *Nat. Methods* **11,** 740–742 (2014).
110. Vallejos, C. A., Marioni, J. C. & Richardson, S. BASiCS: Bayesian Analysis of Single-Cell Sequencing Data. *PLoS Comput. Biol.* **11,** e1004333 (2015).
111. Treutlein, B. *et al.* Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. *Nature* **509,** 371–375 (2014).
112. van der Maaten, L. & Hinton, G. Visualizing Data using t-SNE. *J. Mach. Learn. Res.* **9,**

2579–2605 (2008).

113. Andrews, T. S. & Hemberg, M. Identifying cell populations with scRNASeq. *Mol. Aspects Med.* **59,** 114–122 (2018).
114. Korthauer, K. D. *et al.* A statistical approach for identifying differential distributions in single-cell RNA-seq experiments. *Genome Biol.* **17,** 222 (2016).
115. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36,** 411–420 (2018).
116. Fan, J. *et al.* Characterizing transcriptional heterogeneity through pathway and gene set overdispersion analysis. *Nat. Methods* **13,** 241–244 (2016).
117. Buettner, F., Pratanwanich, N., McCarthy, D. J., Marioni, J. C. & Stegle, O. f-scLVM: Scalable and versatile factor analysis for single-cell RNA-seq. *Genome Biol.* **18,** 212 (2017).
118. Haghverdi, L., Maren, B., Wolf, F. A., Buettner, F. & Theis, F. J. Supplementary Material for : Di ff usion pseudotime robustly reconstructs lineage branching List of Figures. *Nat. Methods* (2016). doi:10.1101/041384
119. Trapnell, C. *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* **32,** 381–386 (2014).
120. Aibar, S. *et al.* SCENIC: Single-cell regulatory network inference and clustering. *Nat. Methods* **14,** 1083–1086 (2017).
121. van Dijk, D. *et al.* Recovering Gene Interactions from Single-Cell Data Using Data Diffusion. *Cell* 1–14 (2018). doi:10.1016/j.cell.2018.05.061
122. Reinius, B. *et al.* Analysis of allelic expression patterns in clonal somatic cells by single-cell RNA-seq. *Nat. Genet.* **48,** 1430–1435 (2016).
123. Shalek, A. K. *et al.* Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature* **498,** 236–240 (2013).
124. Camp, J. G. *et al.* Multilineage communication regulates human liver bud development from pluripotency. *Nature* **546,** 533–538 (2017).
125. Segerstolpe, Å. *et al.* Single-Cell Transcriptome Profiling of Human Pancreatic Islets in Health and Type 2 Diabetes. *Cell Metab.* **24,** 593–607 (2016).
126. Leushacke, M. & Barker, N. Lgr5 and Lgr6 as markers to study adult stem cell roles in self-renewal and cancer. *Oncogene* **31,** 3009–3022 (2012).
127. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* **33,** 495–502 (2015).
128. Karaiskos, N. *et al.* The Drosophila Embryo at Single Cell Transcriptome Resolution.

Science (80-.). **3235,** 117382 (2017).

129. Velten, L. *et al.* Human haematopoietic stem cell lineage commitment is a continuous process. *Nat. Cell Biol.* **19,** 271–281 (2017).
130. Shah, S., Lubeck, E., Zhou, W. & Cai, L. In Situ Transcription Profiling of Single Cells Reveals Spatial Organization of Cells in the Mouse Hippocampus. *Neuron* **92,** 342–357 (2016).
131. Nichterwitz, S. *et al.* Laser capture microscopy coupled with Smart-seq2 (LCM-seq) for robust and efficient transcriptomic profiling of mouse and human cells. *Nat. Commun.* **7,** 1–11 (2016).
132. Janes, K. A., Wang, C. C., Holmberg, K. J., Cabral, K. & Brugge, J. S. Identifying single-cell molecular programs by stochastic profiling. *Nat. Methods* **7,** 311–317 (2010).
133. Ståhl, P. L. *et al.* Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science (80-.).* **353,** 78–82 (2014).
134. Berglund, E. *et al.* Spatial maps of prostate cancer transcriptomes reveal an unexplored landscape of heterogeneity. *Nat. Commun.* **9,** (2018).
135. Wang, X. *et al.* Three-dimensional intact-tissue sequencing of single-cell transcriptional states. **5691,** 1–18 (2018).
136. Lee, J. H. *et al.* Highly multiplexed subcellular RNA sequencing in situ. *Science (80-.).* **343,** 1360–1363 (2014).
137. Hanahan, D. & Weinberg, R. A. Hallmarks of Cancer: The Next Generation. *Cell* **144,** 646–674 (2011).
138. Junttila, M. R. & De Sauvage, F. J. Influence of tumour micro-environment heterogeneity on therapeutic response. *Nature* **501,** 346–354 (2013).
139. Chin, L., Andersen, J. N. & Futreal, P. A. Cancer genomics: From discovery science to personalized medicine. *Nat. Med.* **17,** 297–303 (2011).
140. O'Keefe, B. 'The future of cancer treatment'. *Fortune* **144,** 80 (2001).
141. Malta, T. M. *et al.* Machine Learning Identifies Stemness Features Associated with Oncogenic Dedifferentiation. *Cell* **173,** 338–354.e15 (2018).
142. Navin, N. E. The first five years of single-cell cancer genomics and beyond. *Genome Res.* **25,** 1499–1507 (2015).
143. Burrell, R. A., McGranahan, N., Bartek, J. & Swanton, C. The causes and consequences of genetic heterogeneity in cancer evolution. *Nature* **501,** 338–345 (2013).
144. Ding, L. *et al.* Clonal evolution in relapsed acute myeloid leukaemia revealed by

whole-genome sequencing. *Nature* **481,** 506–510 (2012).

145. Gerlinger, M. *et al.* Intratumor Heterogeneity and Branched Evolution Revealed by Multiregion Sequencing. *N. Engl. J. Med.* **366,** 883–892 (2012).
146. Nik-Zainal, S. *et al.* The life history of 21 breast cancers. *Cell* **149,** 994–1007 (2012).
147. Navin, N. *et al.* Tumour evolution inferred by single-cell sequencing. *Nature* **472,** 90–95 (2011).
148. Kim, C. *et al.* Chemoresistance Evolution in Triple-Negative Breast Cancer Delineated by Single-Cell Sequencing. *Cell* **173,** 879–893.e13 (2018).
149. Casasent, A. K. *et al.* Multiclonal Invasion in Breast Tumors Identified by Topographic Single Cell Sequencing. *Cell* **172,** 205–217.e12 (2018).
150. Fan, J. *et al.* Linking transcriptional and genetic tumor heterogeneity through allele analysis of single-cell RNA-seq data. *Genome Res.* (2018). doi:10.1101/gr.228080.117
151. Wainwright, E. N. & Scaffidi, P. Epigenetics and Cancer Stem Cells: Unleashing, Hijacking, and Restricting Cellular Plasticity. *Trends in Cancer* **3,** 372–386 (2017).
152. Meacham, C. E. & Morrison, S. J. Tumour heterogeneity and cancer cell plasticity. *Nature* **501,** 328–337 (2013).
153. Clevers, H. The intestinal crypt, a prototype stem cell compartment. *Cell* **154,** 274–284 (2013).
154. Doulatov, S., Notta, F., Laurenti, E. & Dick, J. E. Hematopoiesis: A human perspective. *Cell Stem Cell* **10,** 120–136 (2012).
155. Batlle, E. & Clevers, H. Cancer stem cells revisited. *Nat. Med.* **23,** 1124–1134 (2017).
156. Al-Hajj, M., Wicha, M. S., Benito-Hernandez, A., Morrison, S. J. & Clarke, M. F. Prospective identification of tumorigenic breast cancer cells. *Proc. Natl. Acad. Sci.* **100,** 3983–3988 (2003).
157. Chen, J. *et al.* A restricted cell population propagates glioblastoma growth after chemotherapy. *Nature* **488,** 522–526 (2012).
158. Filbin, M. G. *et al.* Developmental and oncogenic programs in H3K27M gliomas dissected by single-cell RNA-seq. *Science (80-.).* **360,** 331–335 (2018).
159. Barker, N. Adult intestinal stem cells: Critical drivers of epithelial homeostasis and regeneration. *Nat. Rev. Mol. Cell Biol.* **15,** 19–33 (2014).
160. Clevers, H. C. & Bevins, C. L. Paneth Cells: Maestros of the Small Intestinal Crypts. *Annu. Rev. Physiol.* **75,** 289–311 (2013).
161. Guinney, J. *et al.* The consensus molecular subtypes of colorectal cancer. *Nat. Med.* **21,** 1350–1356 (2015).

162. Van de Wetering, M. *et al.* The β-catenin/TCF-4 complex imposes a crypt progenitor phenotype on colorectal cancer cells. *Cell* **111,** 241–250 (2002).
163. Brabletz, T. *et al.* Variable Beta-catenin expression in colorectal cancers indicates tumor progression driven by the tumor environment. *Proc. Natl. Acad. Sci.* **98,** 10356–10361 (2001).
164. Schepers, A. G. *et al.* Lineage Tracing Reveals Lgr5+ Stem Cell Activity in Mouse Intestinal Adenomas. *Science (80-.).* **337,** 730–735 (2011).
165. Cortina, C. *et al.* A genome editing approach to study cancer stem cells in human tumors. *EMBO Mol. Med.* **9,** 869–879 (2017).
166. De Sousa E Melo, F. *et al.* A distinct role for Lgr5 + stem cells in primary and metastatic colon cancer. *Nature* **543,** 676–680 (2017).
167. Lambrechts, D. *et al.* Phenotype Moulding of Stromal Cells in the Lung Tumour Microenvironment. *Nature* 1–23 (2018). doi:10.1038/s41591-018-0096-5
168. Patel, A. P. *et al.* Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science (80-.).* **344,** 1396–1401 (2014).
169. Puram, S. V. *et al.* Single-Cell Transcriptomic Analysis of Primary and Metastatic Tumor Ecosystems in Head and Neck Cancer. *Cell* **171,** 1611–1624.e24 (2017).
170. Tammela, T. *et al.* A Wnt-producing niche drives proliferative potential and progression in lung adenocarcinoma. *Nature* **545,** 355–359 (2017).
171. Lim, J. S. *et al.* Intratumoural heterogeneity generated by Notch signalling promotes small-cell lung cancer. *Nature* **545,** 360–364 (2017).
172. Ebinger, S. *et al.* Characterization of Rare, Dormant, and Therapy-Resistant Cells in Acute Lymphoblastic Leukemia. *Cancer Cell* **30,** 849–862 (2016).
173. Azizi, E. *et al.* Single-Cell Map of Diverse Immune Phenotypes in the Breast Tumor Microenvironment. *Cell* (2018). doi:10.1016/j.cell.2018.05.060
174. Sun, Z. *et al.* Single-cell RNA sequencing reveals gene expression signatures of breast cancer-associated endothelial cells. *Oncotarget* **9,** 10945–10961 (2018).
175. Eales, K. L., Hollinshead, K. E. R. & Tennant, D. A. Hypoxia and metabolic adaptation of cancer cells. *Oncogenesis* **5,** e190 (2016).
176. Gilkes, D. M. Implications of hypoxia in breast cancer metastasis to bone. *International Journal of Molecular Sciences* **17,** (2016).
177. Van Den Beucken, T. *et al.* Hypoxia promotes stem cell phenotypes and poor prognosis through epigenetic regulation of DICER. *Nat. Commun.* **5,** 5203 (2014).
178. Klein, C. A. Selection and adaptation during metastatic cancer progression. *Nature* **501,** 365–372 (2013).

179. Cavallo, F., De Giovanni, C., Nanni, P., Forni, G. & Lollini, P. L. 2011: The immune hallmarks of cancer. *Cancer Immunol. Immunother.* **60,** 319–326 (2011).
180. Ye, X. & Weinberg, R. A. Epithelial-Mesenchymal Plasticity: A Central Regulator of Cancer Progression. *Trends Cell Biol.* **25,** 675–686 (2015).
181. Yang, J. *et al.* Twist, a master regulator of morphogenesis, plays an essential role in tumor metastasis. *Cell* **117,** 927–939 (2004).
182. Tran, H. D. *et al.* Transient SNAIL1 expression is necessary for metastatic competence in breast cancer. *Cancer Res.* **74,** 6330–6340 (2014).
183. Creighton, C. J. *et al.* Residual breast cancers after conventional therapy display mesenchymal as well as tumor-initiating features. *Proc. Natl. Acad. Sci.* **106,** 13820–13825 (2009).
184. Kudo-Saito, C., Shirako, H., Takeuchi, T. & Kawakami, Y. Cancer metastasis is accelerated through immunosuppression during epithelial-mesenchymal transition of cancer cells. *Cancer Cell* **15,** 195–206 (2009).
185. Mani, S. a *et al.* The epithelial-mesenchymal transition generates cells with properties of stem cells. *Cell* **133,** 704–715 (2009).
186. Palm, W. & Thompson, C. B. Nutrient acquisition strategies of mammalian cells. *Nature* **546,** 234–242 (2017).
187. Ito, K. & Suda, T. Metabolic requirements for the maintenance of self-renewing stem cells. *Nat. Rev. Mol. Cell Biol.* **15,** 243–256 (2014).
188. Pavlova, N. N. & Thompson, C. B. The Emerging Hallmarks of Cancer Metabolism. *Cell Metab.* **23,** 27–47 (2016).
189. Vander Heiden, M. G. & DeBerardinis, R. J. Understanding the Intersections between Metabolism and Cancer Biology. *Cell* **168,** 657–669 (2017).
190. Nombela-Arrieta, C. *et al.* Quantitative imaging of haematopoietic stem and progenitor cell localization and hypoxic status in the bone marrow microenvironment. *Nat Cell Biol* **15,** 533–43 (2013).
191. Simsek, T. *et al.* The distinct metabolic profile of hematopoietic stem cells reflects their location in a hypoxic niche. *Cell Stem Cell* **7,** 380–390 (2010).
192. Rodríguez-Colman, M. J. *et al.* Interplay between metabolic identities in the intestinal crypt supports stem cell function. *Nature* **543,** 424–427 (2017).
193. Warburg, O. The metabolism of tumours. Investigations from the Kaiser-Wilhelm Institute for Biology, Berlin-Dahlem. *Br. J. Surg.* **19,** 168–168 (1931).
194. Vander Heiden, M. G., Cantley, L. C., Thompson, C. B. & Thompson3, C. B. Understanding the Warburg Effect: The Metabolic Requirements of Cell Proliferation

The Metabolic Requirements of Cell Proliferation. *Science (80-.).* **324,** 1029–1033 (2009).

195. Hensley, C. T. *et al.* Metabolic Heterogeneity in Human Lung Tumors. *Cell* **164,** 681–694 (2016).
196. Lee, M. *et al.* Mathematical modeling links Wnt signaling to emergent patterns of metabolism in colon cancer. *Mol. Syst. Biol.* **13,** 912 (2017).
197. Sancho, P., Barneda, D. & Heeschen, C. Hallmarks of cancer stem cell metabolism. *Br. J. Cancer* **114,** 1305–1312 (2016).
198. Dong, C. *et al.* Loss of FBP1 by snail-mediated repression provides metabolic advantages in basal-like breast cancer. *Cancer Cell* **23,** 316–331 (2013).
199. Song, I. S. *et al.* FOXM1-Induced PRX3 Regulates Stemness and Survival of Colon Cancer Cells via Maintenance of Mitochondrial Function. *Gastroenterology* **149,** 1006–1016 (2015).
200. Sonveaux, P. *et al.* Targeting lactate-fueled respiration selectively kills hypoxic tumor cells in mice. *J. Clin. Invest.* **118,** 3930–3942 (2008).
201. Fessenden, M. Metabolomics: Small molecules, single cells. *Nature* **540,** 153–155 (2016).
202. Damiani, C. *et al.* Integration of single-cell RNA-seq data into metabolic models to characterize tumour cell populations. *Bioinformatics* (2018). doi:10.1101/256644
203. Shaffer, S. M. *et al.* Rare cell variability and drug-induced reprogramming as a mode of cancer drug resistance. *Nature* **546,** 431–435 (2017).
204. Dieter, S. M. *et al.* Distinct types of tumor-initiating cells form human colon cancer tumors and metastases. *Cell Stem Cell* **9,** 357–365 (2011).
205. Dieter, S. M., Glimm, H. & Ball, C. R. Colorectal cancer-initiating cells caught in the act. *EMBO Mol. Med.* **9,** 856–858 (2017).
206. Roerink, S. F. *et al.* Intra-tumour diversification in colorectal cancer at the single-cell level. *Nature* **556,** 437–462 (2018).
207. Mead, B. E. *et al.* Harnessing single-cell genomics to improve the physiological fidelity of organoid- derived cell types. 1–24 (2018). doi:10.1186/s12915-018-0527-2
208. Tirier, S. M. *et al.* pheno-seq - linking 3D phenotypes of clonal tumor spheroids to gene expression. *bioRxiv* 311472 (2018). doi:10.1101/311472
209. Simian, M. & Bissell, M. J. Organoids: A historical perspective of thinking in three dimensions. *J. Cell Biol.* **216,** 31–40 (2017).
210. Djuric, U., Zadeh, G., Aldape, K. & Diamandis, P. Precision histology: how deep learning is poised to revitalize histomorphology for personalized cancer care. *npj*

Precis. Oncol. **1,** 22 (2017).

211. Borten, M. A., Bajikar, S. S., Sasaki, N., Clevers, H. & Janes, K. A. Automated brightfield morphometry of 3D organoid populations by OrganoSeg. *Sci. Rep.* **8,** 5319 (2018).
212. Bithi, S. S. & Vanapalli, S. A. Microfluidic cell isolation technology for drug testing of single tumor cells and their clusters. *Sci. Rep.* **7,** 1–12 (2017).
213. Debnath, J. & Brugge, J. S. Modelling glandular epithelial cancers in three-dimensional cultures. *Nat. Rev. Cancer* **5,** 675–688 (2005).
214. Debnath, J., Muthuswamy, S. K. & Brugge, J. S. Morphogenesis and oncogenesis of MCF-10A mammary epithelial acini grown in three-dimensional basement membrane cultures. *Methods* **30,** 256–268 (2003).
215. Imbalzano, K. M., Tatarkova, I., Imbalzano, A. N. & Nickerson, J. A. Increasingly transformed MCF-10A cells have a progressively tumor-like phenotype in three-dimensional basement membrane culture. *Cancer Cell Int.* **9,** 7 (2009).
216. Soule, H. D. *et al.* Isolation and characterization of a spontaneously immortalized human breast epithelial cell line, MCF-10. *Cancer Res.* **50,** 6075–86 (1990).
217. Dawson, P. J., Wolman, S. R., Tait, L., Heppner, G. H. & Miller, F. R. MCF10AT: a model for the evolution of cancer from proliferative breast disease. *Am. J. Pathol.* **148,** 313–319 (1996).
218. Strickland, L. B., Dawson, P. J., Santner, S. J. & Miller, F. R. Progression of premalignant MCF10AT generates heterogeneous malignant variants with characteristic histologic types and immunohistochemical markers. *Breast Cancer Res. Treat.* **64,** 235–240 (2000).
219. Santner, S. J. *et al.* Malignant MCF10CA1 cell lines derived from premalignant human breast epithelial MCF10AT cells. *Breast Cancer Res. Treat.* **65,** 101–110 (2001).
220. Sommer, C., Straehle, C., Kothe, U. & Hamprecht, F. A. Ilastik: Interactive learning and segmentation toolkit. *Proc. - Int. Symp. Biomed. Imaging* 230–233 (2011). doi:10.1109/ISBI.2011.5872394
221. Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci.* **102,** 15545–15550 (2005).
222. Nieto, M. A., Huang, R. Y. Y. J., Jackson, R. A. A. & Thiery, J. P. P. Emt: 2016. *Cell* **166,** 21–45 (2016).
223. Bach, K. *et al.* Differentiation dynamics of mammary epithelial cells revealed by single-cell RNA sequencing. *Nat. Commun.* **8,** (2017).

224. Gao, R. *et al.* Nanogrid single-nucleus RNA sequencing reveals phenotypic diversity in breast cancer. *Nat. Commun.* **8,** 228 (2017).
225. Attar, M. *et al.* A practical solution for preserving single cells for RNA sequencing. *Sci. Rep.* **8,** 2151 (2018).
226. Giessler, K. M. *et al.* Genetic subclone architecture of tumor clone-initiating cells in colorectal cancer. *J. Exp. Med.* **214,** 2073–2088 (2017).
227. Dalerba, P. *et al.* Single-cell dissection of transcriptional heterogeneity in human colon tumors. *Nat. Biotechnol.* **29,** 1120–1127 (2011).
228. Takebe, N. *et al.* Targeting Notch, Hedgehog, and Wnt pathways in cancer stem cells: Clinical update. *Nat. Rev. Clin. Oncol.* **12,** 445–464 (2015).
229. Sasaki, N. *et al.* Reg4 $^{+}$ deep crypt secretory cells function as epithelial niche for Lgr5 $^{+}$ stem cells in colon. *Proc. Natl. Acad. Sci.* **113,** E5399–E5407 (2016).
230. Bajikar, S. S., Fuchs, C., Roller, A., Theis, F. J. & Janes, K. A. Parameterizing cell-to-cell regulatory heterogeneities via stochastic transcriptional profiles. *Proc. Natl. Acad. Sci.* **111,** E626–E635 (2014).
231. Liberzon, A. *et al.* The Molecular Signatures Database Hallmark Gene Set Collection. *Cell Syst.* **1,** 417–425 (2015).
232. Yan, K. S. *et al.* Intestinal Enteroendocrine Lineage Cells Possess Homeostatic and Injury-Inducible Stem Cell Activity. *Cell Stem Cell* **21,** 78–90.e6 (2017).
233. Wiener, Z. *et al.* Prox1 promotes expansion of the colorectal cancer stem cell population to fuel tumor growth and ischemia resistance. *Cell Rep.* **8,** 1943–1956 (2014).
234. Ragusa, S. *et al.* PROX1 promotes metabolic adaptation and fuels outgrowth of Wnthighmetastatic colon cancer cells. *Cell Rep.* **8,** 1957–1973 (2014).
235. Kuipers, E. J. *et al.* Colorectal cancer. *Nat. Rev. Dis. Prim.* **1,** 1–25 (2015).
236. Muzny, D. M. *et al.* Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487,** 330–337 (2012).
237. Pinto, D., Gregorieff, A., Begthel, H. & Clevers, H. Canonical Wnt signals are essential for homeostasis of the intestinal epithelium. *Genes Dev.* **17,** 1709–1713 (2003).
238. Barker, N. *et al.* Crypt stem cells as the cells-of-origin of intestinal cancer. *Nature* **457,** 608–611 (2009).
239. Merlos-Suárez, A. *et al.* The intestinal stem cell signature identifies colorectal cancer stem cells and predicts disease relapse. *Cell Stem Cell* **8,** 511–524 (2011).
240. Riffle, S. & Hegde, R. S. Modeling tumor cell adaptations to hypoxia in multicellular tumor spheroids. *J. Exp. Clin. Cancer Res.* **36,** 102 (2017).

241. Ziskin, J. L. *et al.* In situ validation of an intestinal stem cell signature in colorectal cancer. *Gut* **62,** 1012–1023 (2012).
242. Eide, P. W., Bruun, J., Lothe, R. A. & Sveen, A. CMScaller: An R package for consensus molecular subtyping of colorectal cancer pre-clinical models. *Sci. Rep.* **7,** 1–8 (2017).
243. Patel, A. P. *et al.* Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science (80-.).* **344,** 1396–1401 (2014).
244. Gabrilovich, D. I. *et al.* H1(0) histone and differentiation of dendritic cells. A molecular target for tumor-derived factors. *J. Leukoc. Biol.* **72,** 285–96 (2002).
245. Torres, C. M. *et al.* The linker histone H1.0 generates epigenetic and functional intratumor heterogeneity. *Science (80-.).* **353,** (2016).
246. Lee, D. D. & Seung, H. S. Learning the parts of objects by non-negative matrix factorization. *Nature* **401,** 788–791 (1999).
247. Wilson, A. *et al.* c-Myc controls the balance between hematopoietic stem cell self-renewal and differentiation. *Genes Dev.* **18,** 2747–2763 (2004).
248. Shyh-chang, N. & Ng, H. The metabolic programming of stem cells. *Genes Dev.* **31,** 336–346 (2017).
249. Bahr, C. *et al.* A Myc enhancer cluster regulates normal and leukaemic haematopoietic stem cell hierarchies. *Nature* **553,** 515–520 (2018).
250. Oliver, P. L. *et al.* Oxr1 Is Essential for Protection against Oxidative Stress-Induced Neurodegeneration. *PLOS Genet.* **7,** e1002338 (2011).
251. Primo-Parmo, S. L., Sorenson, R. C., Treiber, J. & La Du, B. N. The Human Serum Paraoxonase / Arylesterase Gene (PON1) Is One Member of a Multigene Family. *Genomics* **507,** 498–507 (1996).
252. Fan, J. *et al.* Glutamine-driven oxidative phosphorylation is a major ATP source in transformed mammalian cells in both normoxia and hypoxia. *Mol. Syst. Biol.* **9,** (2013).
253. Kaur, R. *et al.* Activation of p21-activated kinase 6 by MAP kinase kinase 6 and p38 MAP kinase. *J. Biol. Chem.* **280,** 3323–3330 (2005).
254. Venteicher, A. S. *et al.* Decoupling genetics, lineages, and microenvironment in IDH-mutant gliomas by single-cell RNA-seq. *Science (80-.).* **355,** eaai8478 (2017).
255. Yu, S. *et al.* Paneth Cell Multipotency Induced by Notch Activation following Injury. *Cell Stem Cell* 1–14 (2018). doi:10.1016/j.stem.2018.05.002
256. Arwert, E. N., Hoste, E. & Watt, F. M. Epithelial stem cells, wound healing and cancer. *Nat. Rev. Cancer* **12,** 170–180 (2012).
257. Khoo, B. L. *et al.* Expansion of patient-derived circulating tumor cells from liquid

biopsies using a CTC microfluidic culture device. *Nat. Protoc.* **13,** 34–58 (2018).

258. Datlinger, P. *et al.* Pooled CRISPR screening with single-cell transcriptome readout. *Nat. Methods* **14,** 297–301 (2017).
259. Huisken, J., Swoger, J., Del Bene, F., Wittbrodt, J. & Stelzer, E. H. K. Optical sectioning deep inside live embryos by selective plane illumination microscopy. *Science (80-.).* **305,** 1007–1009 (2004).
260. Strnad, P. *et al.* Inverted light-sheet microscope for imaging mouse pre-implantation development. *Nat. Methods* **13,** 139–142 (2016).
261. Wu, Y. *et al.* Spatially isotropic four-dimensional imaging with dual-view plane illumination microscopy. *Nat. Biotechnol.* **31,** 1032–1038 (2013).
262. Liu, T.-L. *et al.* Observing the Cell in Its Native State: Imaging Subcellular Dynamics in Multicellular Organisms. *Science (80-.).* **360,** (2018).
263. Argelaguet, R. *et al.* Multi-Omics factor analysis - a framework for unsupervised integration of multi-omic data sets. *bioRxiv* **e8124,** 217554 (2018).
264. Lake, B. B. *et al.* Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain. *Nat. Biotechnol.* **36,** 70–80 (2018).
265. Buenrostro, J. D. *et al.* Integrated Single-Cell Analysis Maps the Continuous Regulatory Landscape of Human Hematopoietic Differentiation. *Cell* **173,** 1535–1548.e16 (2018).
266. Lecun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521,** 436–444 (2015).
267. Pawlowski, N., Caicedo, J. C., Singh, S., Carpenter, A. E. & Storkey, A. Automating Morphological Profiling with Generic Deep Convolutional Networks. *bioRxiv* 4–8 (2016). doi:10.1101/085118
268. Moncada, R. *et al.* Building a tumor atlas: integrating single-cell RNA-Seq data with spatial transcriptomics in pancreatic ductal adenocarcinoma. *bioRxiv* 254375 (2018). doi:10.1101/254375
269. Hoppe, P. S. *et al.* Early myeloid lineage choice is not initiated by random PU.1 to GATA1 protein ratios. *Nature* **535,** 299–302 (2016).
270. Buggenthin, F. *et al.* Prospective identification of hematopoietic lineage choice by deep learning. *Nat. Methods* **14,** 403–406 (2017).
271. Smallwood, S. A. *et al.* Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nat. Methods* **11,** 817–820 (2014).
272. Rotem, A. *et al.* Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state. *Nat. Biotechnol.* **33,** 1165–1172 (2015).
273. Cheow, L. F. *et al.* Single-cell multimodal profiling reveals cellular epigenetic

heterogeneity. *Nat. Methods* **13,** 833–836 (2016).

274. Alemany, A., Florescu, M., Baron, C. S., Peterson-Maduro, J. & Van Oudenaarden, A. Whole-organism clone tracing using single-cell sequencing. *Nature* **556,** 108–112 (2018).
275. Wagner, D. E. *et al.* Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo. *Science (80-.).* **987,** 1–12 (2018).
276. Spanjaard, B. *et al.* Simultaneous lineage tracing and cell-type identification using CrIsPr-Cas9-induced genetic scars. *Nat. Biotechnol.* **36,** 469–473 (2018).
277. Raj, B. *et al.* Simultaneous single-cell profiling of lineages and cell types in the vertebrate brain. *Nat. Biotechnol.* **36,** 442–450 (2018).
278. Adamson, B. *et al.* A Multiplexed Single-Cell CRISPR Screening Platform Enables Systematic Dissection of the Unfolded Protein Response. *Cell* **167,** 1867–1882.e21 (2016).
279. Dixit, A. *et al.* Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. *Cell* **167,** 1853–1866.e17 (2016).
280. Jaitin, D. A. *et al.* Dissecting Immune Circuits by Linking CRISPR-Pooled Screens with Single-Cell RNA-Seq. *Cell* **167,** 1883–1896.e15 (2016).
281. Mezger, A. *et al.* High-throughput chromatin accessibility profiling at single-cell resolution. *bioRxiv* 310284 (2018). doi:10.1101/310284
282. Leone, P. *et al.* MHC class i antigen processing and presenting machinery: Organization, function, and defects in tumor cells. *J. Natl. Cancer Inst.* **105,** 1172–1187 (2013).
283. Bar-Ephraim, Y. E. *et al.* Modelling cancer immunomodulation using epithelial organoid cultures. *bioRxiv* (2018). at <http://biorxiv.org/content/early/2018/08/07/377655.abstract>
284. Janes, K. A. Single-cell states versus single-cell atlases - two classes of heterogeneity that differ in meaning and method. *Curr. Opin. Biotechnol.* **39,** 120–125 (2016).
285. Martinez-jimenez, C. P. *et al.* Aging increase cell to cell transciptional variability upon immune stimulation. *Science (80-.).* **1436,** 1433–1436 (2017).
286. Tetteh, P. W. *et al.* Replacement of Lost Lgr5-Positive Stem Cells through Plasticity of Their Enterocyte-Lineage Daughters. *Cell Stem Cell* **18,** 203–213 (2016).
287. Qiu, X. *et al.* Reversed graph embedding resolves complex single-cell trajectories. *Nat. Methods* **14,** 979–982 (2017).
288. Manno, G. La *et al.* RNA velocity in single cells. *Nature* (2018). doi:https://doi.org/10.1038/s41586-018-0414-6

289. Haghverdi, L., Lun, A. T. L., Morgan, M. D. & Marioni, J. C. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat. Biotechnol.* **36,** 421 (2018).
290. Shaham, U. *et al.* Removal of batch effects using distribution-matching residual networks. *Bioinformatics* **33,** 2539–2546 (2017).
291. Young, M. D. *et al.* Single-cell transcriptomes from human kidneys reveal the cellular identity of renal tumors. *Science (80-.).* **361,** 594–599 (2018).
292. Clark, S. J. *et al.* scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells. *Nat. Commun.* **9,** 781 (2018).
293. Elaimy, A. L. *et al.* VEGF-neuropilin-2 signaling promotes stem-like traits in breast cancer cells by TAZ-mediated repression of the Rac GAP β2-chimaerin. *Sci. Signal.* **11,** (2018).
294. Viale, A. *et al.* Oncogene ablation-resistant pancreatic cancer cells depend on mitochondrial function. *Nature* **514,** 628–632 (2014).
295. Vlashi, E. *et al.* Metabolic state of glioma stem cells and nontumorigenic cells. *Proc. Natl. Acad. Sci.* **108,** 16062–16067 (2011).
296. Pascual G, Avgustinova A, Mejetta S, Martín M, Attolini C, Berenguer A, Toll A, Hueto J, Bescós C, Croce LD, B. S. Targeting metastasis stem cells through the fatty acid receptor CD36. *Nature* **1,** 1–25 (2016).
297. Tang, S. L., Gao, Y. L. & Chen, X. B. Wnt/β-catenin up-regulates midkine expression in glioma cells. *Int. J. Clin. Exp. Med.* **8,** 12644–12649 (2015).
298. Leyns, L., Bouwmeester, T., Kim, S. H., Piccolo, S. & De Robertis, E. M. Frzb-1 is a secreted antagonist of Wnt signaling expressed in the Spemann organizer. *Cell* **88,** 747–756 (1997).
299. A., L. *et al.* Recommendations for reporting tumour budding in colorectal cancer based on the International Tumour Budding Consensus Conference (ITBCC) 2016. *Virchows Arch.* **469,** S172 (2016).
300. Codeluppi, S. *et al.* Spatial organization of the somatosensory cortex revealed by cyclic smFISH. *bioRxiv* **14,** 276097 (2018).
301. Mansour, A. A. *et al.* An in vivo model of functional and vascularized human brain organoids. *Nat. Biotechnol.* **36,** 432–441 (2018).
302. Öhlund, D. *et al.* Distinct populations of inflammatory fibroblasts and myofibroblasts in pancreatic cancer. *J. Exp. Med.* **214,** jem.20162024 (2017).
303. Dijkstra, K. K. *et al.* Generation of Tumor-Reactive T Cells by Co-culture of Peripheral Blood Lymphocytes and Tumor Organoids. *Cell* **0,** (2018).

304. Fatehullah, A., Tan, S. H. & Barker, N. Organoids as an in vitro model of human development and disease. *Nat. Cell Biol.* **18,** 246–254 (2016).
305. Girardot, C., Scholtalbers, J., Sauer, S., Su, S. Y. & Furlong, E. E. M. Je, a versatile suite to handle multiplexed NGS libraries with unique molecular identifiers. *BMC Bioinformatics* **17,** 4–9 (2016).
306. Gu, Z., Eils, R. & Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32,** 2847–2849 (2016).
307. Hochgerner, H. *et al.* STRT-seq-2i: Dual-index 5′ single cell and nucleus RNA-seq on an addressable microwell array. *Sci. Rep.* **7,** 1–8 (2017).
308. Lake, B. B. *et al.* A comparative strategy for single-nucleus and single-cell transcriptomes confirms accuracy in predicted cell-type expression from nuclear RNA. *Sci. Rep.* **7,** 6031 (2017).
309. He, K., Gkioxari, G., Dollar, P. & Girshick, R. Mask R-CNN. *Proc. IEEE Int. Conf. Comput. Vis.* **2017–Octob,** 2980–2988 (2017).
310. Lin, T. Y. *et al.* Microsoft COCO: Common objects in context. *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)* **8693 LNCS,** 740–755 (2014).

5 Materials & Methods

The following Material and Methods section (especially section 5.1) is widely adapted from the pheno-seq manuscript[208] for which I have written the original text including the associated methods part.

5.1 Pheno-seq – linking morphological and functional features to gene expression in 3D culture systems

5.1.1 Breast cancer model MCF10CA

5.1.1.1 Cell culture

The cell line MCF10CA1d clone 1 (acquired from The Barbara Ann Karmanos Cancer Institute), a transformed derivative of the immortalized breast epithelial cell line MCF10A, was cultured and passaged in 25 cm^2 culture flasks (greiner bio-one). Cells grown in 2D were cultured in growth medium composed of DMEM/F12 medium supplemented with 5% horse serum, 10 μg/ml Insulin (Life Technologies), 20 ng/ml EGF, 0.5 mg/ml hydrocortisone and 100 ng/ml Cholera Toxin (Sigma). Cells were passaged with 0.05% Trypsin (Life Technologies) at 80% confluency.

For 3D 'on top' culture, cells were grown in assay medium (growth medium with 2% horse serum and 5 ng/ml EGF) in 24-well cell culture plates (greiner bio-one). As a reconstituted basement membrane, a bed of laminin-rich hydrogel (Matrigel, Corning) was generated by pipetting 70 μm cold Matrigel into the center of pre-wetted 24-well plates. The Matrigel bed was dried for 20 min at 37°C before adding cells. For seeding single cells on top of the Matrigel bed, 2D cultures were dissociated to single-cell suspensions, washed in assay medium, passed through a 35 μm cell strainer and counted by a LUNA automated cell counter (Logos Biosystems). Finally, 4000 cells in 400 μl assay medium (supplemented with 5% Matrigel) were seeded per well by adding cell suspensions in a 45° angle to the wall of the well, resulting in a uniform distribution of cells throughout the well. Assay medium was replaced every 3 days and cells cultured for up to 12 days. In general, all experiments were carried out after 5 days culture on Matrigel.

5.1.1.2 Spheroid recovery from hydrogel

After culturing MCF10CA cells in 3D for 5 days, medium was removed completely and 500 μl filtered and pre-warmed Dispase (Sigma) was added. The Matrigel bed was then detached from the wells by scratching over the well bottom with a 1000 μl pipette tip and the

suspension was resuspended carefully five times. Spheroids were then incubated at 37°C for 7 min, transferred to a 15 ml falcon and 5 ml assay medium was added and resuspended slowly with a 5 ml pipette. Afterwards, spheroids were spun down (300 g, 3 min) and resuspended in DMEM (Life Technologies). At these steps, it is not recommended to use PBS due to perturbation of the spheroid morphology. Generally, this procedure resulted in approximately 2000 isolated spheroids per well.

5.1.1.3 Spheroid isolation and dissociation to single-cell suspensions

For isolation and classification of individual MCF10CA spheroids before dissociation, spheroid suspensions were diluted in assay medium to 100 spheroids per ml and distributed into GravityTRAP ultra-low attachment 96-well plates in 10 µl per well (PerkinElmer). Plates were then centrifuged for 2 min at 250 g. V-shaped and 1 mm flat-bottom wells enabled efficient classification of spheroids (round vs. aberrant) with an inverted brightfield microscope and 10x or 20x objectives. After isolation and identification of 30-40 spheroids of each class, 50 µl Accumax was added to each well and spheroids were dissociated to single cell for 10-15 min at 37°C. Shear forces were applied to stimulate dissociation by resuspending wells of one class with the same 200 µl pipette. After a second incubation time of 5 min at 37°C, wells of one class were pooled in 1.5 ml microcentrifuge tubes, centrifuged at 300 g for 3 min and resuspended in either assay medium or DMEM/F12.

5.1.1.4 Reseeding assay

For independent seeding of cells derived from round and aberrant spheroid phenotypes, spheroids were isolated, dissociated and pooled as described above (section 5.1.1.3). During dissociation, a 10 µl Matrigel bed was generated in 15µ angiogenesis slides (Ibidi). Cells were resuspended in 50 µl assay medium (supplemented with 5% Matrigel) and added to pre-treated angiogenesis slides. Medium was replaced every 3 days and cells were cultured for up to 6 days.

5.1.1.5 Single-cell capture, mRNA library preparation and sequencing

For preparation and sequencing of single-cell RNA sequencing libraries, spheroids were dissociated as described above (section 5.1.1.3) and resuspended in DMEM/F12 medium. Capture, lysis, full-length cDNA synthesis and amplification was performed on the Fluidigm C1 Single-Cell Auto Prep IFC. Cell suspensions at a concentration of 350 cells/µl were mixed with C1 Cell Suspension Reagent (Fluidigm; ratio of 4:1) before loading on the IFC. Single-cell capture was checked with an inverted brightfield microscope and capture sites

with only 1 cell were marked. Protocol and reagents for single-cell RNA extraction, reverse transcription (RT) and mRNA amplification amplification (21 amplification cycles) were used as written in the SMARTer Ultra Low RNA Kit for Fluidigm C1. Sequencing libraries were generated with the Nextera XT kit (Illumina) according to the adapted Fluidigm protocol. Concentration, quality and size of cDNA and Nextera XT libraries was assessed with a fluorometer (Qubit) and by on-chip electrophoresis (Agilent Bioanalyzer high sensitivity DNA chips). Libraries of up to 24 cells were pooled and sequenced on an Illumina HiSeq 2000 machine (1 × 50-bp single-end reads).

5.1.1.6 Manual pheno-seq workflow, library preparation and sequencing

For RNA-sequencing of libraries derived from manually isolated single spheroids (manual pheno-seq), spheroid suspensions were diluted to 500/ml and 2 µl was carefully dispensed to the wall of the well of GravityTRAP 96-well ULA plates. After vertical tapping of the plate, wells with single spheroids were identified and classified with an inverted brightfield microscope. Prior to RT and amplification, RNA was extracted with the Arcturus PicoPure kit (ThermoFisher). Therefore, 50 µl extraction buffer was directly added to 96-wells, incubated for 2 min at room temperature and transferred to 1.5 ml microcentrifuge tubes. RNA was extracted as described in the PicoPure Kit (Appendix B and Section 4B.2) including DNase digestion (Appendix A, RNase-Free DNAse Set, Qiagen). RNA was eluted in Nuclease-free water (~10 µl) which was used as input for full-length cDNA synthesis and amplification amplification (16 amplification cycles) by the SMART-Seq v4 Ultra Low Input RNA Kit for sequencing (TakaraBio). Sequencing libraries were generated with the Nextera XT kit (Illumina) as described in the SMART-Seq v4 protocol. Concentration, quality and size of cDNA and Nextera XT libraries was assessed with a fluorometer (Qubit) and by on-chip electrophoresis (Agilent Bioanalyzer high sensitivity DNA chips). Ten libraries were pooled and sequenced on an Illumina HiSeq 2000 machine (1 × 50-bp single-end reads).

5.1.1.7 High-throughput pheno-seq workflow, library preparation and sequencing

The nanowell-based iCELL8 scRNA-seq system (TakaraBio), that integrates imaging and gene expression profiling of big samples of up to 100 µm[83], was adapted and improved for high-throughput (HT-)pheno-seq. For fluorescence detection of cytoplasm and DNA, spheroids were first stained for three hours with 10 µM CellTracker Red CMTPX dye and 1 µg/ml Hoechst 33258 (ThermoFisher). To acquire a high number of spheroids for dispensing into the 5,184 nanowell chip, spheroids from 6-8 wells were recovered as described above (section 5.1.1.2) and washed with 7 ml DMEM (Life Technologies). Three

wells were combined per 15 ml falcon tube for centrifugation. The reversible cross-linker dithio-bis(succinimidyl) propionate (DSP) was prepared for cellular fixation as described previously[225] in PBS and directly filtered through a 10 µm strainer before usage (PluriSelect) to avoid excessive precipitation. Spheroids were then resuspended in 400 µl DSP at a working concentration of 1 mg/ml and incubated at room temperature for 30 min. Spheroids were then washed two times with cold PBS (centrifugation at 650 g and 500 g, 3 min, 4°C) and finally resuspended in 650 µl cold PBS with 1x second diluent (for iCELL8) and 0.4 U/µl recombinant RNase Inhibitor (TakaraBio). Spheroids were dispensed into the barcoded 5184-nanowell chip (version 1) with the iCELL8 Single-Cell System (TakaraBio) as described in the Rapid Development Protocol (in-chip RT-PCR amplification). As a control, we first processed one chip without cellular fixation using the default chip setup, the standard microscope and provided CellSelect software.

The following modifications were applied for improved HT-pheno-seq: Wells in the 384-well source plate were mixed with a 200 µl pipette tip between dispensing intervals in order to minimize spheroid settling in source plate wells for enabling even distribution of spheroids in nanowells. Next, the iCELL8 chip was tightly sealed with a strongly adhesive imaging foil (TakaraBio) similar to the standard single-cell protocol. However, instead of spinning cells to the bottom, spheroids were centrifuged upside-down to the foil (700 g, 5 min, 4°C) in order to reduce the objective working distance and to avoid light reflections inside the well during imaging. We used an inverted confocal laser-scanning microscope (Leica SP8) with 10x objective (2x2 wells per field of view) instead of the standard and system-integrated fluorescence wide-field microscope with 4x objective (6x6 wells per field of view) to further enhance imaging resolution. After imaging, spheroids were centrifuged to the well bottom (700 g, 5 min, 4°C) and chips were frozen and stored at -80°C.

A KNIME image pre-processing workflow as well as the PhenoSelect software were used for spheroid detection and interactive selection (for more detailed description of microscopy, image pre-processing and PhenoSelect section 5.1.3.5). The ‘filter file’ generated by PhenoSelect was used to dispense reagents in selected nanowells as described in the Rapid Development Protocol (TakaraBio), with the exception that we adjusted the amount of Triton-X100 to a final well concentration of 1% for enhanced efficiency of spheroids lysis (Master mix: 52.8 µl 5 M Betaine, 24 µl 25 mM dNTP mix (TakaraBio), 3.2 µl 1 M $MgCl_2$ (Invitrogen), 8.8 µl 100 mM Dithiothreitol (TakaraBio), 61.9 µl 5x SMARTScribe first-strand buffer, 33.3 µl 2x SeqAmp PCR buffer, 4.0 µl 100 µM RT E5 Oligo, 8.8 µl 10 µM Amp primer (all TakaraBio), 4.8 µl 100% Triton X-100 (Acros), 28.8 µl SMARTScribe Reverse Transcriptase, 9.6 µl SeqAmp DNA Polymerase (TakaraBio)). The maximum spheroid size

(that correlates with the number of cells per spheroid/well) should not exceed 100-150 μm to avoid any negative influence on RT efficiency. Furthermore, lysis reagents, concentration and timing might have to be adjusted for different 3D cell culture models.

After in-chip mRNA reverse transcription and cDNA amplification (18 amplification cycles, In-chip RT/Amp Rapid Development Protocol) inside a modified SmartChip Cycler (Bio-Rad), libraries were pooled, concentrated (DNA Clean and Concentrator–5 kit, Zymo Research) and purified using 0.6x Ampure XP beads. Barcoded cDNA was processed to 3´-end sequencing libraries by the Illumina Nextera XT kit with adaptions described in the Rapid Development Protocol. Concentration, quality and size of cDNA and Nextera XT libraries was assessed with a fluorometer (Qubit) and by on-chip electrophoresis (Agilent Bioanalyzer high sensitivity DNA chips). Improved HT-pheno-seq paired-end iCELL8 libraries (21 + 70 bp paired-end reads) were sequenced on an Illumina NextSeq 500 machine in high-output mode. The 'bottom control' chip without improved imaging was sequenced on a HiSeq 2000 machine with similar settings. However, the 'bottom control' was only used to assess library quality and not for further downstream analysis.

5.1.2 Colon TICs spheroids

5.1.2.1 Cell culture

The primary patient-derived colon tumor spheroid culture derived from a liver metastasis was established as described previously[204]. Human CRC tissue was obtained from Heidelberg University Hospital in accordance with the declaration of Helsinki. Informed consent on tissue collection was acquired from each patient and approved by the University Ethics Review Board. CRC cells were cultured in 75 cm^2 ultra-low attachment flasks (Corning) in advanced D-MEM/F-12 medium supplemented with 0.6% Glucose, 2 mM L-glutamine (Life Technologies), 4 µg/ml heparin, 5 mM HEPES, 4 mg/ml BSA (Sigma), 10 ng/ml FGF basic and 20 ng/ml EGF (R&D Systems). EGF/FGF was added every 4 days and medium was exchanged every 4-8 days. For dissociation, spheroid cultures were spun down for 5 min at 900 rpm and resuspended in 2-4 ml 0.25% Trypsin (Life Technologies). Shear forces were applied with a 1000 μl pipette every 5 min for 20 min in total to stimulate dissociation. Subsequently, 4-8 ml stop solution (PBS supplemented with 20% heat inactivated and sterile filtered fetal bovine serum, Life Technologies) was added and cells were spun down for 5 min at 900 rpm. For passaging, cells were resuspended in medium, passed through a 40 μm strainer and counted.

5.1.2.2 Reseeding assay

We cultured colon spheroids for 10 days and performed a stepwise size exclusion by (reverse-) filtering with standard 100 µm, 70 µm, 40 µm and 20 µm cell strainers in order to independently isolate, dissociate and reseed cells from different size classes (70-100 µm and 20-40 µm). Spheroids were dissociated to single-cell suspension as described above (section 5.1.2.1) with the exception that cells were passed through a 15 µm cell strainer. Finally, 50,000 cells were re-seeded in 60 mm Ultra Low Attachment Culture Dishes (Corning). Growth factors were added every 4 days and cells cultured for 10 days. Culture dishes were shaken every day to avoid clustering of spheroids.

5.1.2.3 γ-secretase inhibitor assay

To selectively inhibit the γ-secretase machinery during growth of spheroids, the γ-secretase inhibitor PF-03084014 (Sigma) was dissolved in sterile and distilled water (stock concentration 1 mM). Cells were dissociated as described above and 20,000 cells were seeded in 24-well ultra-low attachment plates (Corning) in the presence of PF-03084014 at final concentrations of 5, 10 or 20 µM. In addition, we included a solvent control for the maximum amount of added water (20 µl). Growth factors were added every 4 days and cells were cultured for 10 days.

5.1.2.4 Single-cell culture and HT-pheno-seq of colon tumor spheroids

To specifically culture single floating CRC spheroids, spheroids were dissociated, passed through a 15 µm cell strainer and counted as described above (section 5.1.2.1). Cells were seeded in Aggrewell400 6-Well plates (StemCell Technologies) in which each well contains an array of approximately 7000 inverse pyramidal shaped microwells with a size of 400 x 400 µm. Wells were pre-treated according to the manufacturer´s instructions and washed once with PBS and once with medium. Afterwards, 3500 cells in 3 ml medium were added in a 45° angle to the wall of the well, which resulted in uniform distribution of cells in microwells. EGF/FGF was added every 4 days and cells were cultured for 10 days, resulting in 300-400 spheroids (>20 µm) per 6-well. Spheroids isolated from 4-6 plates (24-36 6-wells, 168,000-252,000 microwells) were stained for three hours with 10 µM CellTracker Red CMTPX dye and 1 µg/ml Hoechst 33258 (ThermoFisher), harvested, pooled and washed once with FluoroBrite DMEM (Life Technologies, 900 rpm for 5 min).

HT-pheno-seq was performed as described for MCF10CA spheroids above (section 5.1.1.7) but with following modifications: In contrast to MCF10CA spheroids, colon spheroids were not fixed by DSP because spheroid isolation did not involve contact loss from Matrigel. To

minimize disassembly of spheroids during processing, cells were dispensed in FluoroBrite DMEM instead of PBS.

5.1.3 Microscopy and image analysis

5.1.3.1 Image processing and analysis

Acquired microscopy images were processed and analyzed using KNIME Image Processing (https://www.knime.com/community/image-processing, Version 3.2.1), ImageJ/FIJI (https://imagej.nih.gov/ij/), R (Version 3.3.1)/R studio (https://www.rstudio.com/) and/or Graph Pad Prism 7 (https://www.graphpad.com /scientific-software/prism/). The ggplot2 package implemented in R and Graph Pad Prism 7 were used for visualization of data and the PhenoSelect web app is based on the shiny package (https://shiny.rstudio.com).

5.1.3.2 Assessing single-cell seeding efficiency

Wells with single MCF10CA (section 5.1.1.1) or CRC cells (section 5.1.2.4) stained with Hoechst 33258 (1 μg/ml) and CellTracker Red CMPTX (10 μM) were imaged one hour after seeding with a confocal laser-scanning microscope (Leica SP8) equipped with a 10x/0.30 air objective (Leica HC PL FLUOTAR). Images of cells in 24 wells (MCF10CA) or 6-well AggreWell400 (CRC spheroids) were pre-processed and analyzed with custom made KNIME image analysis workflows (Triplicates, three independent wells). Based on the CellTracker signal, images were first flattened by Gaussian convolution (sigma=1) followed by otsu's method for global thresholding. Next, water shedding (default ImageJ settings) and a minimum filter (minimum 5 pixels) were applied. Cells touching the border were excluded from further analysis. To detect doublets, the same workflow was used for a duplicated set of images, but with radially increasing the size of the labels (Max filter node, span 6) after thresholding, resulting in larger cell masks. Generated labels were then projected back on the primary generated original single-cell segmentations. All cell masks with more than two single-cell segmentations in enlarged segmentations were counted as doublets.

5.1.3.3 Reseeding assay

For reseeded cells derived from isolated spheroid phenotypes of MCF10CA (section 5.1.1.4) and CRC cultures (5.1.1.4), images were acquired with a Zeiss LSM780 Axio Observer confocal laser scanning microscope equipped with a 10x/0.3 air objective (Zeiss EC PLAN-NEOFLUAR) in brightfield.

For MCF10CA spheroids, a training dataset was first generated based on randomly seeded and cultured cells, whereas classification was based on independently reseeded cells from

round and aberrant 3D phenotypes. The training dataset was used for the 'Pixel classification' option in the random-forest machine learning software ilastik[220]. Spheroids and background pixel classes were labeled with the paintbrush tool and assigned as 'round' and 'aberrant' and iterative training allowed the generation of stable probability maps to distinguish the three object types 'round', 'aberrant' and 'background'. For automated analysis of MCF10CA reseeding assays, a custom KNIME workflow loaded the previously trained project file using the 'ilastik headless node' and images to be classified were imported into KNIME and classified by applying the trained model to each image. The probability maps for 'round' and 'aberrant' spheroids were smoothed using a manual threshold of 0.5 and a size threshold (3000 pixels) was applied for all objects based on the probability maps. Spheroids of both phenotype classes were automatically quantified and assigned to their respective experiment and condition.

For CRC cells derived from size classes ´big´ (70-100 μm) and ´small´ (20-40 μm), 8 x 8 images per well were automatically acquired in 6-well plates (Greiner) using a custom Zeiss VBA macro. Images were analyzed using a custom KNIME workflow: Briefly, edges were detected using the default function "Find Edges" (implemented in ImageJ). Subsequently, images were thresholded manually (value=20) and default watershedding algorithm was applied to each image to separate neighboring spheroids. After segmentation, only segments of 500 to 1x10^5 pixels (1 pixel = 0.73 µm) were used and counted for downstream analysis.

5.1.3.4 γ-secretase inhibitor assay

Spheroids treated with the γ-secretase inhibitor PF-03084014 that have been cultured for 10 days (section 5.1.2.3) were stained with CellTracker Red CMPTX (10 μM) for 3 hours before imaging with a Leica SP8 confocal laser-scanning microscope equipped with a 10x/0.30 air objective (Leica HC PL FLUOTAR). 5 x 5 images per well (10 Z-stacks per position) were acquired automatically using the 'TileScan' option to directly stitch acquired images of one well to one final composite image.

Stiched CellTracker Red images were then analyzed using a custom KNIME workflow: Briefly, acquired Z-stacks were merged using minimum intensity projection and the local contrast was enhanced by ´Contrast Limited Adaptive Histogram Equalization´ (CLAHE) in order to correct for unevenness of wells (8 contextual regions, 256 bins, slope=6.0). Afterwards, images were median-filtered, manually thresholded (value=20) and connected component analysis was applied to assign labels to objects/spheroids. Only segments > 200

pixels (1 pixel = 0.44 µm) were used for downstream analysis to exclude single-cells. Spheroid size (area) was calculated in KNIME using the ´Feature Calculator´ node.

5.1.3.5 HT-pheno-Seq microscopy, image processing and PhenoSelect

For image acquisition of all 5,184-nanowells after dispensing of spheroids (sections 5.1.1.7 and 5.1.2.4), chips were fixed on a metallic Chip Spinner (TakaraBio) with conventional adhesive tape and placed into a standard plate holder for inverted imaging. Nanowells were imaged upside-down automatically with an inverted Leica SP8 confocal laser scanning microscope system: To span four nanowells per field of view, we used a 10x/0.30 air objective (Leica HC PL FLUOTAR) but images were acquired with 0.9x digital zoom. Excitation was set to 405 and 552 nm and emission filters were set to receive signals between 415 – 485 nm (for DNA/Hoechst) and 555 – 625 nm (for cytoplasm/CellTracker Red). Laser intensity and gain were slightly optimized for every experiment, but the pinhole aperture was set to 5.0 Airy Units permanently. To extrapolate the correct focus position for each well, the ´predictive focus´ option was used. Images had a resolution of 512 x 512 pixels, with 2.53 µm/pixel. A pre-defined HCS A template of the LAS X microscope software (Leica) was used for grid design matching the 72 x 72 nanowell chip dimensions. Scanning of one chip with the abovementioned settings took approximately 30 minutes, thereby resulting in 2 x 1296 images.

The first part of the image analysis workflow in KNIME/ImageJ was used for image pre-processing, including assignment of correct well positions, cropping, detection and segmentation, as well as feature extraction and quantification: Image names were first changed to match order of image acquisition and well location. The resulting format is comparable to the in-build iCELL8 microscope, although with 2 x 2 instead of 6 x 6 wells per field of view. Importantly, images were rotated 90° to correct for the camera orientation of the Leica SP8 system. Next, image cropping generated images containing only one nanowell per image. Based on the segmentation of the field of view containing the four most round segments, a mask was built for each well over all positions for cropping of images containing four nanowells. Since the SP8 microscope can image four wells so that only minimal offset between well positioning was visible, this method allowed cropping of all nanowells over all positions. Next, names of images with single wells were transformed to row/column positions to exactly match iCELL8-specific barcode assignments. As an example, the four wells of the first top-left image were transformed to ‘0_0’, ‘0_1’, ‘1_0’ and ‘1_1’. Segmentation, feature extraction and analyses were only performed on the cytoplasmic signal (CellTracker Red). The image was first smoothed by a Gaussian convolution algorithm (sigma = 5), then

manually thresholded (value=20) and spheroids in close proximity were separated by the watershed algorithm (ImageJ command 'watershed', default parameters). Only objects between 25 and 40,000 pixels were considered for downstream analysis to exclude imaging artefacts and single-cells. Spheroids touching the border of images were excluded. Cropped nanowell images were saved automatically and individually for each channel but also as overlay of the two channels for better visualization in the shiny web app. Finally, a .csv file was generated containing names and calculated 2D image features (derived from KNIME image processing node 'Feature Calculator') of nanowells containing at least one spheroid.
The second part of the image analysis workflow in a shiny web app format (PhenoSelect, Supplementary Figure 4) has been developed for interactive analysis and selection of spheroids for sequencing: A saved .csv file containing quantified spheroid features is automatically handled by a custom R script and directly embedded together with the images in an interactive R/shiny application (PhenoSelect). Thereby, images and visualization of associated image feature statistics can be browsed interactively. In addition, direct visual inspection of given image features over the whole population allows the identification of distinct subtypes, e.g. by a particular shape or size. For quantification of absolute spheroid sizes, the respective spheroid major axis length value (in pixels) was multiplied with the physical length of a pixel in segmented objects. Spheroids can be selected based on applied thresholds and/or and individual wells can be selected/discarded manually. A list of selected wells can be saved and reloaded to proceed with selection at later time-points. Once spheroids had been selected for sequencing, PhenoSelect generated a 'filter file', which is used to instruct the iCELL8 dispenser software. In addition, a ´well-list file´ was generated containing well-barcode assignments for demultiplexing and calculated image features for selected spheroids. Finally, we implemented visualization of image features in t-SNE maps based on gene expression computed by PAGODA. After sequencing and analysis of selected spheroids, this tab enabled combined analysis for direct association of functional visual phenotypes to transcriptomic heterogeneity.

5.1.3.6 Leakage test

To assess potential leakage between wells, we dispensed a highly fluorescent solution of PBS + 1 µg/ml fluorescein sodium salt (Sigma) into one half of the nanowells and dispensed only PBS into the other half and into control wells. A specific dispensing pattern was chosen to generate a high number of borders between nanowells filled with fluorescein and those filled with PBS. Afterwards, the chip was handled and imaged as described above (section 5.1.1.7) but with laser and filter sets matching fluorescent properties of fluorescein (λ_{ex} 460

nm; λ_{em} 515 nm). Acquired images were processed for well assignment, segmentation and cropping by the HT-pheno-seq pre-processing workflow and average fluorescence intensity was measured for every well. Average fluorescence intensity values ranging from 0 to 255 (8 bit) were color coded and plotted onto a 72 x 72 grid resembling the iCELL8 chip layout using a custom R script.

5.1.3.7 Antibody staining for immunofluorescence

Whole mount immunofluorescence staining of MCF10CA spheroids was performed as described previously[214]. Briefly, cells were fixed in 24-wells with 2% Methanol-free Formaldehyde solution (ThermoFisher) for 20 min at room temperature and washed 2x with PBS. Spheroids were permeabilized with PBS + 0.5% TritonX-100 (Sigma) for 10 min and washed 3x with PBS + 75 mg/ml Glycine (pH=7.4, Sigma). Unspecific binding sites were blocked for one hour at room temperature with 10% goat serum in IF-wash solution (PBS + 5 mg/ml NaN_3, 10 mg/ml bovine serum albumin, 2% TritonX-100 and 0.4% Tween20, pH=7.4, Sigma). Subsequently, primary antibodies (in blocking solution) were added and incubated at 4°C overnight. The next day, cells were washed three times with IF-wash and then incubated with fluorescently labeled secondary antibodies in blocking solution for one hour at room temperature if primary antibodies were unlabeled. Afterwards, cells were washed three times with IF-wash and two times with PBS and then incubated in PBS + 1 µg/ml Hoechst for 20 min at room temperature. Cells were then washed with PBS, removed from the surface with a 1,000 µl pipette and transferred into 8-well Nunc Lab-Tek Chamber Slides (ThermoFisher) for improved fluorescence detection. Following antibodies were used in this study: Rabbit anti-Vimentin antibody Alexa Fluor® 594 (1:100, EPR3776, abcam), mouse anti-β-Actin antibody (1:200, 8H10D10, Cell Signaling), Mouse anti-Cytokeratin 15 antibody (1:50, LHK15, ThermoFisher), Goat anti-mouse Alexa Fluor 594 (1:200, Cell Signaling).

3 x 3 images per well (20 Z-stacks per position) were acquired automatically on a Zeiss LSM780 Axio Observer confocal microscope equipped with a 10x/0.3 air objective (Zeiss EC PLAN-NEOFLUAR) using a custom Zeiss VBA macro. Lasers and filters were set to measure fluorescence emitted from Hoechst (DNA) and from Alexa Fluor 594-labeled antibodies. Brightfield images were obtained in parallel.

Acquired images were analyzed using a custom KNIME workflow as follows: Z-stacks were first merged by average intensity projection and masks for single spheroids were created based on the Hoechst signal. To generate spheroid masks, images were smoothed by Gaussian convolution (sigma=2) and Otsu´s method was used for thresholding. Labels were assigned to objects by connected component analysis and objects < 300 and > 800,000

pixels were filtered out to remove segmentation artifacts. For comparison of expression of antibody targets between round and aberrant 3D phenotypes, single spheroids were manually classified as ‘round’ or ‘aberrant’ based on brightfield images. Protein abundances were defined as mean pixel intensity of the fluorescence signal emitted from labeled antibodies per spheroid.

5.1.3.8 RNA FISH

For histological preparation of MCF10CA spheroids, cells were cultured and isolated as described above (section 5.1.1.1), fixed with 2% Formaldehyde solution for 20 min at room temperature and washed 2x with PBS. Next, spheroids were incubated in PBS + 15% sucrose (Sigma) and PBS + 30% sucrose for cryopreservation (both 15 min at room temperature), embedded in Richard-Allan Scientific™ Neg-50™ Frozen Section Medium (ThermoFisher) and frozen in the gaseous phase of liquid nitrogen.

For histological preparation of CRC spheroids, different size classes of spheroids (> 70 μm and 20-40 μm) derived from single-cells were isolated by (reverse-) filtering (sections 5.1.2.2 and 5.1.2.4). The filtering step was added in order to distinguish between small spheroids and big spheroids that were cut in peripheral regions. Spheroids were fixed with 4% Formaldehyde solution for 20 min at 4°C, washed 2x with PBS and incubated in 30% sucrose for cryopreservation at 4°C overnight. The next day, spheroids were embedded in Neg-50™ and frozen in the gaseous phase of liquid nitrogen.

Sectioning was performed at –20°C on a cryostat (Leica) and 10 μm slices were mounted on Superfrost Plus slides (ThermoFisher). Specimens and cryosections were stored at -80°C until further use.

For RNA fluorescence in-situ hybridization (RNA-FISH), we used the RNAscope Fluorescent Multiplex Assay 2.0 (ACDbio). Cryosections were processed as described in the ‘Sample Preparation Technical Note for Fixed Frozen Tissue’ and the ‘Fluorescent Multiplex Kit User Manual PART 2’. Briefly, cryosections were pre-treated with Protease IV (ACDbio) for 15 min at room temperature. Subsequently, mRNA-specific probes were hybridized at 40°C for 120 min followed by stepwise hybridization of probes for signal amplification and fluorescent detection (Amp-1-FL – Amp-4-FL). Up to three transcripts were labeled by Alexa488, Atto550 and Atto647 fluorescent dyes (overview of used RNA-FISh probes in Supplementary Table 10). Cryosections were counterstained with DAPI, mounted in SlowFade Gold Antifade solution (ThermoFisher) and stored at 4°C until further use.

RNA-FISH microscopy was performed with a Leica SP8 confocal laser-scanning microscope equipped with a 40x/1.30 oil objective (Leica HC APO CS2). Individual spheroids were

imaged at 1024 x 1024 pixel-resolution semi-automatically using the 'Mark and Find' option in the Leica SP8 acquisition software. To acquire signals from the whole 10 μm cryosection height, a Z-range of 20 μm was imaged by 15 stacks (1.43 μm distance between frames). Lasers and filters were set to match fluorescent properties of DAPI and RNA-FISH dyes. MCF10CA spheroids were classified as 'aberrant' or 'round' manually during imaging, whereas CRC spheroid classes were separated during sample preparation.

For analysis of RNA-FISH images we used a custom KNIME workflow: Z-stacks were merged using maximum intensity projection and masks for single spheroids were created using the DAPI signal. Acquired DAPI signals were smoothed by Gaussian convolution (sigma = 5) and a maximum filter with a radius of 12 pixels was applied, resulting in individual masks for all spheroids within one image. Only the biggest spheroid per image was used for analysis if two or more objects were present in one field of view. To approximate transcript abundances (measured as fluorescence intensities derived from specifically labeled probes) we first corrected for background noise by fitting two local maxima (j and k) to the pixel intensity histogram of each spheroid using the 'intermodes' function in KNIME. Based on the determined pixel intensity threshold between two maxima calculated as (j+k)/2 (probe-specific), we defined the relative transcript expression per spheroid as quantified pixel % that exceeds this threshold per spheroid.

5.1.3.9 Cell count determination by light sheet imaging and 3D segmentation

To approximate cell numbers from images acquired in CRC HT-pheno-seq experiments, we generated a 3D image reference dataset to calculate the linear relationship of size (area) and cell numbers. CRC spheroids were stained for three hours with CellTracker Red CMPTX (10 μM) and 1 μg/ml Hoechst and subsequently isolated and fixed as described above (section 5.1.3.8). Next, spheroids were mounted in 2% low-melting agarose (Sigma) and image stacks were acquired using a Dual-View Inverted Selective Plane Illumination Microscope (ASI di-SPIM) using Nikon 40x/0.80W NA NIR-Apo water dipping objectives. Dual-view raw image data was processed to generate isotropic images at a resolution of 0.325px/μm (400 images per Z-stack, 0.325 μm distance). Image pre-processing and 3D segmentation was performed with a custom KNIME workflow as follows: A 2D projection of the smoothed CellTracker image was used to produce a mask that was generated for each spheroid individually. Obtained 2D masks covering spheroids were applied to all individual slices in order to count nuclei only within this area and exclude artifacts. Single cells were first segmented and counted for each slice individually. Next, overlapping cells were separated by watershedding (KNIME node: Waehlby Cell Clump Splitter). Afterwards, the 2D

segments were used as ‘seeds’ for 3D Voronoi segmentation in order to obtain a 3D segmentation that accounts for cells spanning multiple slices. Cell numbers and corresponding 2D spheroid dimensions were exported from KNIME and analyzed using a custom R script. First, we converted the respective pixel numbers for minor and major axis of all 2D spheroid masks to metric distances. To account for variable spheroid morphologies, we further estimate the spheroid size as the product of its biggest and smallest diameter (minor and major axis). Afterwards, we plotted the measured size of every spheroid against its respective cell count determined by 3D segmentation in KNIME. A linear model was fitted through the acquired data points and the obtained slope was used to calculate cell count approximations for spheroids o HT-pheno-seq experiments.

5.1.4 Sequencing data analysis

5.1.4.1 Pre-processing of RNA-seq data and library quality control

For single-cell and spheroid RNA-seq data pre-processing, an automated in-house workflow based on Roddy (https://github.com/TheRoddyWMS/Roddy) was established. Briefly, read quality was evaluated using FastQC. For iCELL8 libraries, barcodes from the 1st 21 bp read were assigned to the associated nanowell with the Je demultiplexing suite[305]. Remaining primer sequences, Poly-A/T tails and low-quality ends (<25) were trimmed using Cutadapt. Furthermore, since NextSeq (Illumina) encodes undetected bases as incorrect ‘Gs’ with high quality, Cutadapt’s ‘—nextseq-trim’ option was used for improved quality trimming. Trimmed reads were mapped to the reference genome hs37d5 (derived from the 1000 genomes project) using the STAR aligner. Mapped BAM files were quantified using featureCounts with reference annotation gencode v19.

RNA-seq libraries were filtered out that did not match the following criteria: MCF10CA scRNA-seq: (i) > 300,000 reads, (ii) > 3000 detected genes (i.e. > 0 read count), (iii) < 10% mitochondrial reads; MCF10CA pheno-seq: (i) > 100,000 reads, (ii) > 2000 detected genes, (iii) < 15% mitochondrial reads; Colon spheroid pheno-seq: (i) > 200,000 reads, (ii) > 3000 detected genes, (iii) < 15% mitochondrial reads.

For performance comparison of scRNA-seq and pheno-seq methods in detecting genes, MCF10CA sequencing libraries were downsampled to 100,000 reads with a custom R script.

Spheroids of HT-pheno-seq datasets with imaging artifacts (e.g. segmentation errors) were removed if detected during combined downstream analysis.

5.1.4.2 RNA-seq subpopulation and differential expression analysis

We analyzed transcriptional heterogeneity by pathway and gene set overdispersion analysis (PAGODA/SCDE-package[116]) to identify expression signatures that separate distinct cellular subpopulations. Genes with less than 10 mapped reads in the whole dataset were excluded from further analysis. First, PAGODA generates error models for cells/spheroids using a binominal/Poisson mixture model, thereby controlling for technical aspects of variability, like effective sequencing depth, drop-out rate and amplification noise, respectively. For K-nearest neighbor error modelling, k was set to 30 (for manual pheno-seq and pseudo pheno-seq dataset: k=3), and the minimum number of reads required to be considered non-failed was set to 2. Subsequently, PAGODA performs weighted principal component analysis (wPCA) on *de-novo* identified and annotated gene sets in order to identify gene sets that exhibit statistically significant variability. Generally, the scores for the first principal component (PC) are presented. Annotated hallmark (H) and gene ontology (GO_C5) gene sets were downloaded from the Molecular Signature Database (MSigDB). *De-novo* gene sets were identified by hierarchical clustering (Ward method; dendrogram was cut into 150 clusters). Gene set overdispersion was calculated as Z-score relative to the genome-wide model and corrected Z-scores (cZ) were computed using multiple hypothesis testing by the Holm procedure. Hierarchical clustering was then performed on top significant aspects of heterogeneity and redundant aspects were grouped with a similarity threshold of 0.7. Up to ten top significant aspects were used for visualization. In addition, 2D t-SNE maps[112] were generated based on PAGODA's weighted Pearson correlation distances. In addition, the following confounding expression signatures (e.g. technical aspect or cell cycle influence) were removed using the 'pagoda.substract.aspect' function:

1.) All datasets were corrected for the influence of gene coverage (estimated as a number of genes with non-zero magnitude per cell)

2.) MCF10CA scRNA-seq: GO_REGULATION_OF_CELL_CYCLE and HALLMARK _G2M_CHECKPOINT;

3.) MCF10CA HT-pheno-seq: GO_ NUCLEOSIDE_MONOPHOSPHATE_ METABOLIC _PROCESS, GO_MITOCHONDRIAL_ENVELOPE, GO_STRUCTURAL _MOLECULE _ACTIVITY, GO_ HOMEOSTATIC_PROCESS and corresponsing *de-novo* identified gene sets.

Differentially expressed genes (MCF10CA: fold change > 1.3; adjusted p-value < 0.1; CRC: fold change > 1.5; adjusted p-value < 0.05) between detected subpopulations that refer to observed visual phenotypes (k-means clustering, k=2) were identified by the SCDE-package[109].

5.1.4.3 In-silico reconstruction of pseudo pheno-seq profiles from single-cell data

MCF10CA pseudo-spheroid RNA-seq profiles were constructed from scRNA-seq data by randomly dividing cells either derived from round or aberrant phenotypes in four groups each. Read counts for each gene were then averaged over each group, resulting in eight pseudo-spheroid profiles (4 round and 4 aberrant) that were then analyzed by PAGODA similar to the manual pheno-seq dataset. Calculations were carried out in four independent randomizations.

5.1.4.4 Deconvolution of the CRC spheroid dataset by maximum likelihood inference

To infer heterogeneous gene regulatory states informative for single cell expression by deconvolution, we adapted the maximum likelihood inference approach previously developed to identify cell-cell heterogeneities from random 10-cell samples[230] (Stochastic Profiling). In contrast to the previous version of the algorithm, we allowed each sample to consist of different numbers of cells (implemented in the R package stochprofML version 2.0: https://github.com/fuchslab/stochprofML)

As the algorithm assumes that the expression of a spheroid linearly scales with its cell number, we approximated absolute counts per spheroid by using estimated cell numbers derived from light sheet microscopy and image analysis (section 5.1.3.9): First, counts per spheroid were divided by the respective estimated cell number and the minimal average mRNA count per cell over the whole dataset was determined (2374.644). Subsequently, we downsampled the whole dataset to 2300 counts per cell resulting in a perfect correlation of mRNA counts and cell numbers. The downsampled dataset was then filtered by removing genes with less than one count per well on average over the original CRC HT-pheno-seq dataset and genes were removed with less than 5 counts in at least two wells, leaving 13,868 genes that have been taken into account during the profiling procedure. To avoid problems with zeros and log-normal distributions, all zeros were transformed to 0.1.

5.1.4.5 Statistical analysis and visualization

Statistical analysis and visualization of sequencing data was performed in R (Version 3.3.1) or R studio (https://www.rstudio.com/) using PAGODA/SCDE, ggplot2, ComplexHeatmaps[306], the stats package (R version 3.3.1), stochprofML (R version 3.4.1) and in Graph Pad Prism 7 (https://www.graphpad.com/scientific-software/prism/). Gene set enrichment analysis was done by calculating overlaps between identified signatures and

gene sets derived from the Molecular Signature Database[221] (MSigDB, https://software.broadinstitute.org/gsea/ msigdb).

5.1.5 Data and code availability

Raw sequencing data files for MCF10CA datasets are accessible at the European Nucleotide Archive (https://www.ebi.ac.uk/ena) under Accession Number PRJEB26737.
CRC HT-pheno-seq raw sequencing data are accessible at the European Genome - Phenome Archive (http://www.ebi.ac.uk/ega/) under Accession Number EGAS00001002999.

KNIME image analysis workflows, R code for PhenoSelect and PAGODA/SCDE RNA-seq analysis as well as a download link for MCF10CA HT-pheno-seq image data and associated components to run the pre-processing workflow and/or the PhenoSelect web app for interactive selection of spheroids can be found in the pheno-seq github repository (https://github.com/eilslabs/pheno-seq). More detailed information on the automated in-house RNA-seq workflow is available upon request. The newest version of stochProfML 3.4.1 can be found under: https://github.com/lisaamrhein/stochprofML.

5.2 Heterogeneous metabolic signatures are linked to cancer cell differentiation in a 3D model of colorectal cancer

5.2.1 Cell culture and staining

Primary patient-derived colon tumor sphere cultures were received from Heidelberg University Hospital and established as described previously[204]. From each patient, informed consent on the collection of tissue was obtained and approved by the Ethics Board of the University. Cells were cultured in 75 cm^2 ultra-low attachment flasks in advanced D-MEM/F-12 medium supplemented with Glucose (0.6%), 2 mM L-glutamine (Life Technologies), 4 µg/ml heparin, 5 mM HEPES, 4 mg/ml BSA (Sigma), 10 ng/ml FGF basic and 20 ng/ml EGF (R&D Systems). Growth factors were added every 4 days and medium was exchanged every 4-8 days. For passaging, sphere cultures were centrifuged for 5 min at 900 rpm and resuspended in 2-4 ml 0.25% Trypsin (Life Technologies). Cells were trypsinized for 10-20 min depending on the culture. Subsequently, 4-8 ml stop solution (PBS + 20% heat inactivated and sterile filtered fetal bovine serum, Life Technologies) was added and cells were centrifuged for 5 min at 900-1000 rpm. Cells were then resuspended in medium, passed through a 40 μm strainer and counted. For long-term storage, 2-4 million cells were frozen in medium supplemented with 30% fetal bovine serum and 15% dimethyl sulfoxide (DMSO, Roth). For mitochondrial staining, MitoTracker Red CMXRos (Thermo Fisher) was added to spheroids at a final concentration of 100 nM for three hours.

5.2.2 Preparation of single cell suspensions for single cell RNA-sequencing

For scRNA-seq experiments, cells were cultured for 6-14 days after trypsinization depending on the growth rate of the respective patient culture (Supplementary Table 3). Transcriptional heterogeneity induced by cell culture artifacts like hypoxia in inner cores were avoided by limiting the diameter of spheroids to ~150 μm, thereby ensuring general oxygen supply. For dissociation to single cell suspensions, sphere cultures were centrifuged for 5 min at 900 rpm and resuspended in 4 ml 0.25% Trypsin (Life Technologies). To stimulate dissociation, shear forces were applied with a 1000 μl pipette every 5 min for 10-30 min in total depending on the culture (Supplementary Table 3). Subsequently, 8 ml stop solution (PBS + 20% heat inactivated and sterile filtered fetal bovine serum, Life Technologies) was added and cells were centrifuged for 5 min at 1000 rpm. Cells were washed twice in PBS (room-temperature) passed through a 15-20 μm cell strainer (PluriSelect) and counted by a LUNA automated cell counter (Logos Biosystems).

5.2.3 Nanogrid based single cell library preparation and RNA sequencing

The TakaraBio iCELL8 system and the associated Rapid Development Protocol (in-chip RT-PCR amplification) were used for single cell isolation, reverse transcription and cDNA amplification[83]. Single cell suspensions were stained with Hoechst and Propidium Iodide (ReadyProbe Cell Viability Imaging Kit, Invitrogen) for 10 min at room temperature and cell numbers and viability was checked with a Countess automated cell counter (Thermo Fisher). Samples were discarded if cell viability was below 85%. Cell suspensions were diluted to 25 cells/μl in a 384-well source plate and distributed into a barcoded nanowell chip (oligo dT primer with unique barcode for every well) with a multi-sample microsolenoid valve dispenser, thereby achieving up to 30% nanowells with single cells due to Poisson distribution. All libraries were generated using chip version 1 except for Patient 8 (chip version 2: randomized distribution of barcodes in nanowell chip, 14 bp instead of 11 bp well barcode). Subsequently, the nanowell chip is sealed and centrifuged at 300g for 5 min and wells were imaged using an automated fluorescent microscope. After imaging, chips were frozen at -80°C until further use. Images were processed and analyzed using the CellSelect software and manually curated in order to exclude non-detected doublets or dead cells. A filter file generated by the CellSelect software was used to instruct the dispenser to deposit reagents for reverse transcription and amplification only into selected wells. After thawing frozen chips, 50 nl of RT/Amp solution was dispensed into selected nanowells (Master mix: 56 μl 5 M Betaine, 24 μl 25 mM dNTP mix (TakaraBio), 3.2 μl 1 M $MgCl_2$ (Invitrogen), 8.8 μl 100 mM Dithiothreitol (TakaraBio), 61.9 μl 5x SMARTScribe™ first-strand buffer, 33.3 μl 2x SeqAmp™ PCR buffer, 4.0 μl 100 μM RT E5 Oligo, 8.8 μl 10 μM Amp primer (all TakaraBio), 1.6 μl 100% Triton X-100 (Acros), 28.8 μl SMARTScribe™ Reverse Transcriptase, 9.6 μl SeqAmp™ DNA Polymerase (TakaraBio)). After in-chip RT/Amp amplification (18 amplification cycles, in-chip RT/Amp Rapid Development protocol) inside a modified SmartChip Cycler (Bio-Rad), libraries were pooled, concentrated (DNA Clean and Concentrator–5 kit, Zymo Research) and purified using 0.6x Ampure XP beads. Concentration and quality of cDNA was assessed by a fluorometer (Qubit) and by electrophoresis (Agilent Bioanalyzer high sensitivity DNA chips). Next generation sequencing libraries were constructed using the Nextera XT kit (Illumina) following the manufacturer´s instructions. Final libraries were sequenced with the NextSeq 500 system in high-output mode (paired-end, 21 x 70 for v1, 24 x 67 for v2 chip).

5.2.4 scRNA-seq data analysis

5.2.4.1 Pre-processing of RNA-seq data, library quality control and normalization

For pre-processing of CRC single-cell RNA-seq data, an automated in-house workflow based on Roddy was used. Read quality was evaluated using FastQC. iCELL8 library barcodes from the first 21 bp read (24 bp for chip version 2) were assigned to the associated nanowell with the Je demultiplexing suite[305]. Remaining primer sequences, Poly-A/T tails and low-quality ends (<25) were trimmed using Cutadapt. Furthermore, since NextSeq (Illumina) encodes undetected bases as incorrect 'Gs' with high quality, Cutadapt's '—nextseq-trim' option was used for improved quality trimming. Trimmed reads were mapped to the reference genome hs37d5 (derived from the 1000 genomes project) using the STAR aligner. Mapped BAM files were quantified using featureCounts with reference annotation gencode v19.

RNA-seq libraries were filtered out that did not match the following criteria: (i) > 100,000 reads, (ii) > 1000 detected genes, (iii) < 15% mitochondrial reads. In addition, we removed strong PCA outliers (section 5.2.4.2) and libraries with top 5% of reads for every patient independently. The latter was done in order to control for non-detected cell doublets[307].

Adapting a previously published approach[169], we quantified expression levels based on raw read counts as $E_{i,j} = \log_2(CPM_{i,j} / 10 + 1)$, with $CPM_{i,j}$ referring to count-per-million for gene *i* in sample *j*. To focus on genes expressed at high or intermediate levels, we calculated the aggregate expression of each gene across all cells as $E_a = E_{i,j} = \log_2(\text{mean}[E_{i,1...n}] + 1)$, and excluded genes with $E_a < 3.5$.

5.2.4.2 Analysis of inter-tumor heterogeneity and subtype classification

Filtered and normalized data of all patients combined was used to characterize inter-patient differences in gene expression. For identification of highly variable genes, principal component analysis (PCA), clustering and differential expression analysis, we used the Seurat package[81] as implemented in R.

Highly variable genes were identified using the 'FindVariableGenes' function (with y.cutoff = 0.5), and PCA was used for dimensionality reduction prior to clustering. The number of significant PCs was determined using the 'PCElbowPlot' function, and cells were clustered using the 'FindClusters' function (with resolution = 1.0) on the significant PCs only. tSNE[112] was used to visualize clustering results and patient-specific marker genes were identified by differential expression analysis using the Wilcoxon Rank Sum test. Expression levels of the top 10 differentially expressed genes from every patient were clustered based on average

group linkage (UPGMA) of Pearson correlation coefficients, using the 'aheatmap' function from the 'NMF' package in R.

Consensus Molecular Subtypes (CMS) of CRC as defined by the Cancer Genome Atlas[161] were assigned for each patient independently by 'CMScaller' implemented in R[242], either for each cell individually or after pooling data from all cells, with the false discovery rate set to 0.25.

5.2.4.3 Analysis of intra-tumor heterogeneity to identify shared expression programs

To correct for global inter-patient shifts in gene expression levels, we first calculated relative expression levels for every patient independently by mean-centering: $Er_{i,j} = E_{i,j} - \text{mean}[E_{i,1...n}]$ and then combined the data of all patients.

As a first approach, for each patient separately, we applied PCA to highly variable genes (as implemented in Seurat (section 5.2.4.2), and genes that were identified to be highly correlated with significant PCs were evaluated for biological relevance. The top 30 genes with high and low PC scores in the first principal component were clustered based on average group linkage (UPGMA) of Pearson correlation coefficients, using the 'aheatmap' function from the 'NMF' package in R.

As an alternative, we adapted an approach based on non-negative matrix factorization (NNMF[246]) previously used for the analysis of scRNA-seq data derived from primary and metastatic head and neck cancers[169] to more precisely identify variable gene expression programs (meta-signatures) that are shared between patients. In contrast to the previously applied approach, we did not apply NNMF to every patient individually but used mean-centered data of all patients combined. We applied NNMF as implemented in MATLAB ('nnmf' function) to the mean-centered data from all LGR5$^+$ CRC cultures (Table 2), with the number of factors set to k=25 and all negative values (including drop-out events) set to zero. As drop-outs, defined as genes that are expressed but not detected due to technical reasons, are known to be problematic for scRNA-seq data normalization and analysis, using NNMF is most likely beneficial as it inherently discounts their effects.

In order to exclude signatures that are specific to one or a few patients, we calculated pairwise overlaps in the frequency distributions across cells of factor scores for individual factors and excluded factors that did not show a 50% overlap in at least five patients. Next, factors were analyzed for biological relevance by GSEA[221] and by manual evaluation, and factors likely to be driven by technical artifacts were excluded. Subsequently, we defined meta-signatures as the averaged expression of the top 200 genes per factor and merged redundant meta-signatures that showed similar enrichments and clusterings by combining

genes from different factors (core-signatures). Meta-signature scores were clustered using complete linkage of Euclidean distances. We repeated NNMF analysis with various numbers of factors which resulted in the identification of similar core signatures (not shown).

To assess the degree to which individual cells express specific gene expression programs, we adapted a previously described cell scoring approach based on the expression of pre-defined meta-signatures[169] that uses control random gene sets as a background model in order to control for technical confounders such as library complexity. In addition, we binarized the meta-signature scores for each cell by defining a meta-signature as 'ON' if its expression was more than one standard deviation above the mean across all cells, and 'OFF' otherwise.

For refinement of signature scores to compare lineage-specific cell states for cell cycle, OXPHOS, hypoxia and glycolysis (Figure 2.16), we extracted genes from cell state meta-signatures that actually overlap with HALLMARK gene sets for functional enrichments of interest (MSigDB[231]). The following HALLMARK gene sets were used: HALLMARK_OXIDATIVE_PHOSPHORYLATION (OXPHOS); HALLMARK_HYPOXIA (hypoxia); HALLMARK_GLYCOLYSIS (glycolysis); HALLMARK_G2/M, HALLMARK_MITOTIC_SPINDLE, HALLMARK_DNA_REPAIR (cell cycle). Similar to meta-signature scores, cell state scores were defined as averaged expression per gene set.

5.2.5 *In-situ* analysis of gene expression by microscopy

5.2.5.1 Histological preparation and multiplexed RNA-FISH

For histological preparation, spheroids of Patients 1, 4 and 5 were fixed with 4% Formaldehyde solution for 20 min at 4°C, washed twice with PBS and incubated in 30% sucrose at 4°C overnight. The next day, spheroids were embedded in Neg-50™ and frozen in the gaseous phase of liquid nitrogen.

Histological sectioning was performed at –20°C on a cryostat (Leica) and 10 μm slices were mounted on Superfrost Plus slides (ThermoFisher). Embedded specimens and cryosections were stored at -80°C until further use.

For RNA fluorescence in-situ hybridization (RNA-FISH), we used the RNAscope® Fluorescent Multiplex Assay 2.0 (ACDbio). Cryosections were processed as described in the 'Sample Preparation Technical Note for Fixed Frozen Tissue' and the 'Fluorescent Multiplex Kit User Manual PART 2'. Cryosections were first pretreated with Protease IV (ACDbio) for 15 min at room temperature. Next, transcript-specific probes were hybridized at 40°C for 120 min followed by stepwise hybridization of probes for signal amplification and fluorescent detection (Amp-1-FL – Amp-4-FL). Up to three transcripts were labeled by Alexa488,

Atto550 and Atto647 fluorescent dyes (overview of used RNA-FISH probes in Supplementary Table 10). For parallel mitochondrial and mRNA staining, the Atto550 probe has been replaced by the Mitotracker Red CMXRos dye (section 5.2.1). Finally, cryosections were counterstained with DAPI, mounted in SlowFade Gold Antifade solution (ThermoFisher) and stored at 4°C until further use.

5.2.5.2 RNA-FISH microscopy

RNA-FISH images were acquired on a Leica SP8 confocal laser-scanning microscope equipped with a 40x/1.30 oil objective (Leica HC APO CS2). Images of individual spheroids at pixel resolution 1024 x 1024 (16 bit) were generated semi-automatically using the 'Mark and Find' option in the Leica SP8 acquisition software. To cover the 10 μm cryosection height by imaging, a Z-range of 20 μm was acquired by 15 stacks (1.43 μm distance between frames). Lasers and filters were set to match the fluorescent properties of DAPI and abovementioned dyes/probes.

5.2.5.3 RNA-FISH image analysis

For quantitative analysis of RNA-FISH imaging data, we estimated cellular transcript in cells by measuring fluorescence signals of mRNA targeting probes in single nuclei. Although we thereby miss cytoplasmic signal, relative mRNA abundances between nuclei and whole cells have been shown to be highly correlated[308].

To prepare spheroid images for further analysis, we performed Maximum Intensity Projection on each channel separately. For automated nuclei instance detection and segmentation in spheroids, a deep learning object detection and instance segmentation workflow incorporating Mask R-CNN[309] was implemented. The neural network was initialized using pre-trained models trained on the Microsoft COCO: Common Objects in Context dataset[310] and fine-tuned using images of nuclei acquired from various unrelated sources. The maximum intensity projections of the DAPI images were used as inputs for the neural network to produce segmentation for each individual nucleus as outputs. The nuclei sizes were calculated using these segmented DAPI masks, and objects < 350 pixels were filtered out and excluded from subsequent analysis.

For analyzing transcript abundance, maximum intensity projections of the RNA-FISH channels were binarized using the 'Maximum Entropy' thresholding method in FIJI/ImageJ. Transcript abundance was estimated by overlaying the nuclei masks on the maximum projected probe channels and calculating the number of pixels that lies within each mask. For quantification of mitochondrial abundance per cell, MitoTracker signals were binarized

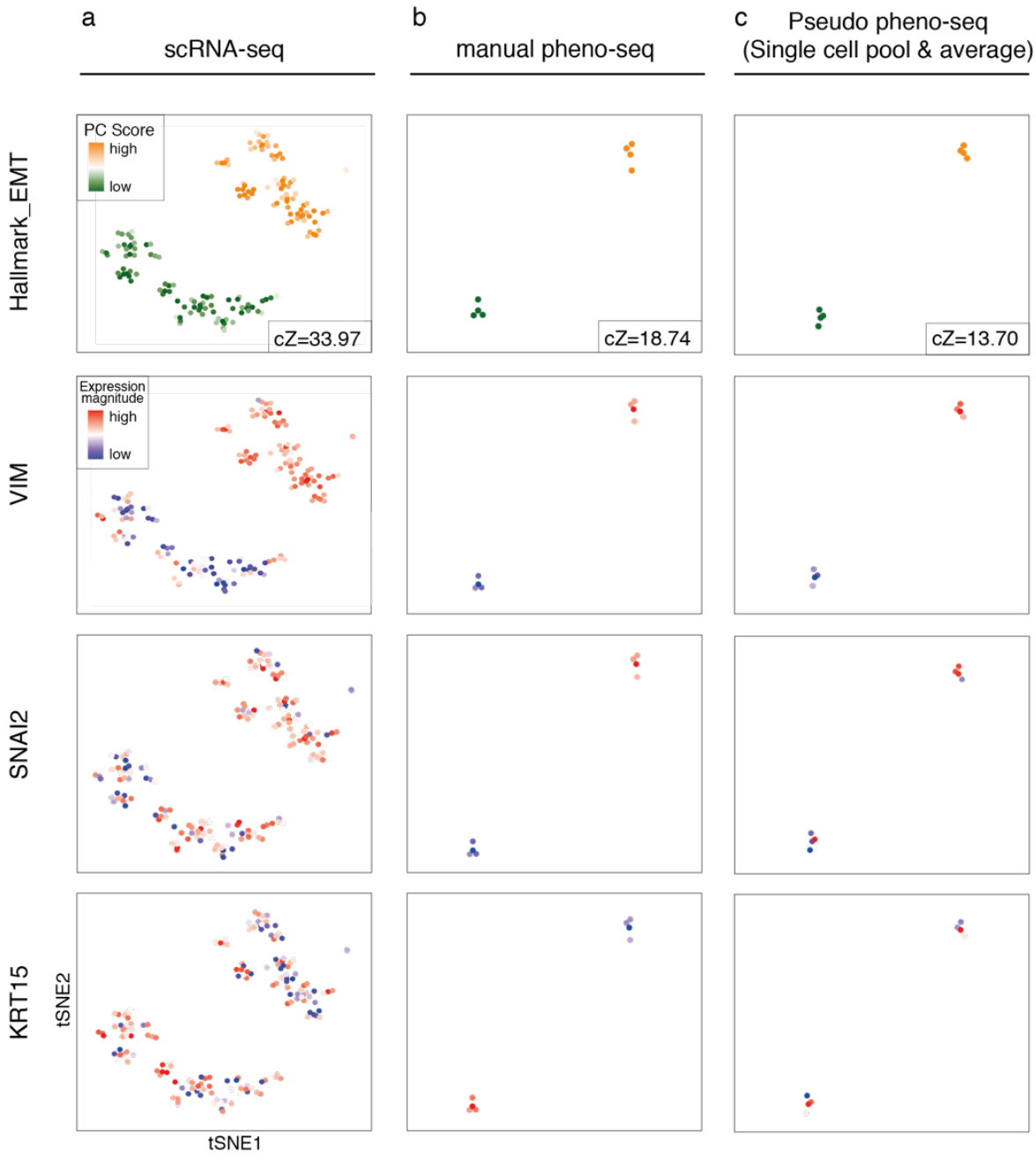
using the 'Moments' thresholding method in FIJI/ImageJ. In order to account for the MitoTracker signals that are predominantly localized outside of the nuclei masks, we expanded the nuclei before pixel counting by morphological dilation (two iterations) as implemented in scikit-image (Python). For RNA-FISH and Mitotracker signals, we then performed k-means clustering on frequency distributions of pixel counts per cell (Nucleus) to identify and separate the cells into two distinct 'ON' (high expression/abundance) and 'OFF' (low expression/abundance) states. k = 2 was used for LGR5 and DEFA5 mRNA probes, while k = 3 was used for MKI67 and Mitotracker signals to better capture gradual differences between cells. Venn diagrams were computed with the matplotlib-venn package as implemented in Python.

5.2.6 Statistical analysis

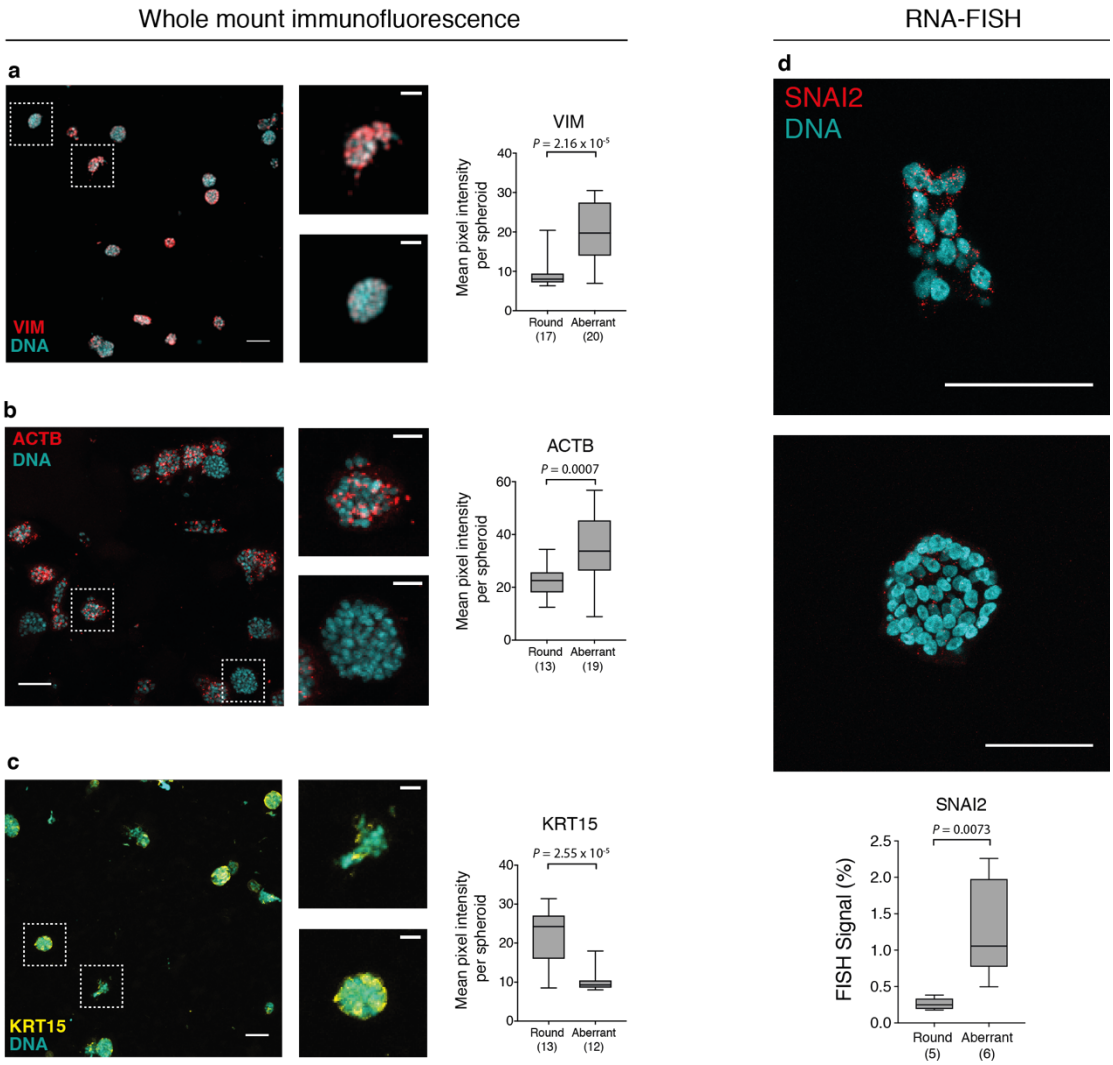
Statistical analysis and visualization of sequencing data was performed in R (Version 3.3.1) or R studio (https://www.rstudio.com/) using Seurat, ggplot2, the stats package (R version 3.3.1), in MATLAB using non-negative matrix factorization ('nnmf') and in Graph Pad Prism 7. Gene set enrichment analysis was done by calculating overlaps between identified signatures and gene sets derived from the Molecular Signature Database[221] (MSigDB, https://software.broadinstitute.org/gsea/ msigdb). Image pre-processing, statistical analysis and visualization of was performed in Python, R studio and in Graph Pad Prism 7.

6 Appendix

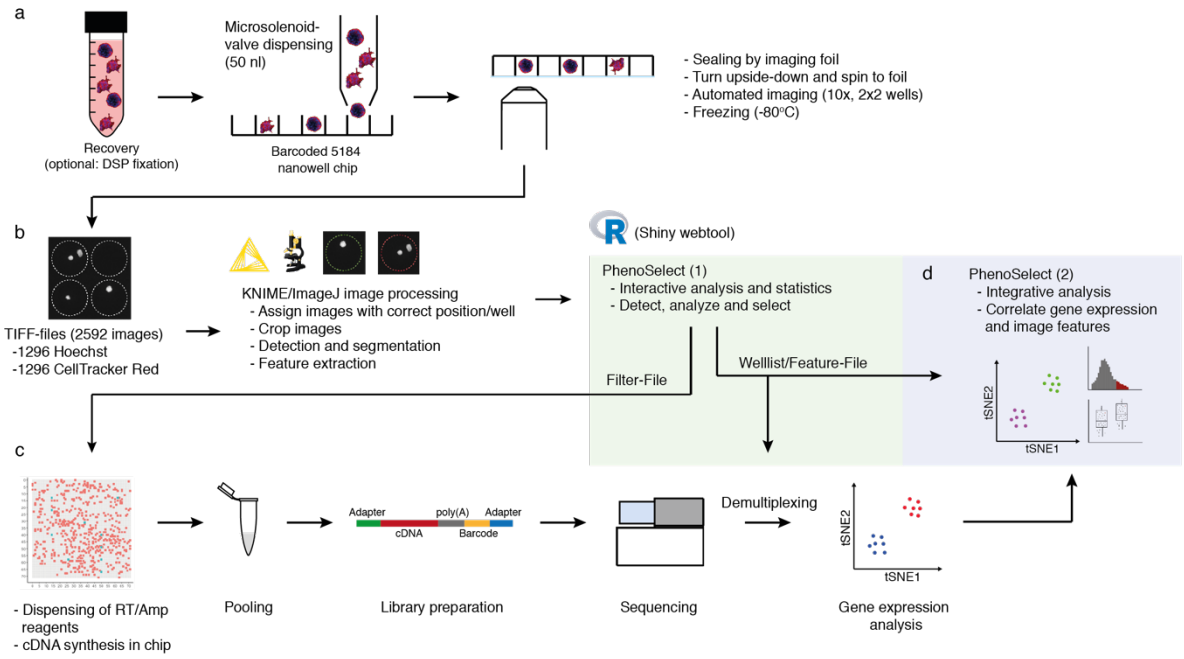
6.1 Supplementary Figures



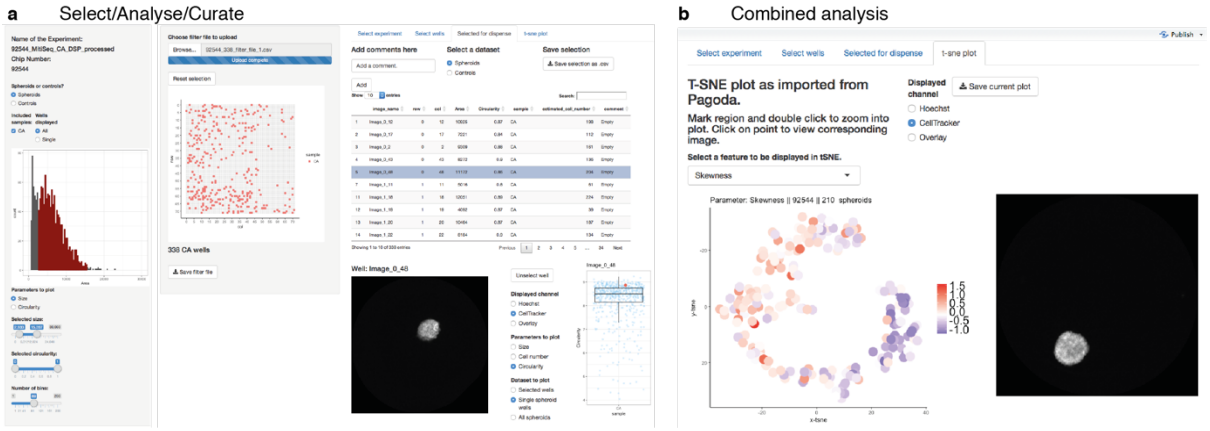
Supplementary Figure 1 | Comparison of scRNA-seq and manual pheno-seq by tSNE visualizations. (a, b, c) PAGODA tSNE visualizations of MCF10CA scRNA-seq (a), manual pheno-seq (b) and Pseudo pheno-seq (c, based on averaged single cell data, first randomization also shown in Figure 2.3). Datasets are colored by PC scores for HALLMARK_EMT gene sets (including associated cZ scores as measure of gene set overdispersion) and by expression magnitude of spheroid phenotype-specific markers VIM, SNAI2 and KRT15. RNA-seq analysis has been performed together with Jeongbin Park, Simon Steiger and Zuguang Gu.



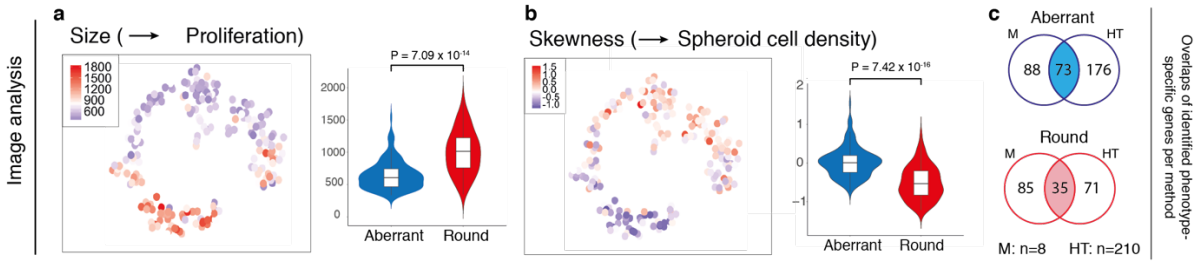
Supplementary Figure 2 | Validation of pheno-seq data by quantitative fluorescence microscopy. (a, b, c) IF staining with primary antibodies targeting VIM (a), ACTB (b) and KRT15 (c). Images show Z-projections of whole mount spheroid immunofluorescence images. Plotted values represent mean pixel intensity per classified spheroid of the respective phenotype class. Dashed boxes in overview images (scale bar 100 μm) correspond to magnified images (scale bar 30 μm). **(d)** RNA-FISH with probe targeting SNAI2 mRNA (Alexa488, scale bar 100 μm). Images represent Z-projections. Plotted values represent pixel percentage that exceeds the threshold per spheroid of the respective class after background correction. **(a-d)** All samples are counterstained with Hoechst dye for nuclei visualization (Hoechst: cyan; antibody signal for 'round' specific markers: yellow; labelled antibodies and RNA-FISH probe for 'aberrant' specific markers: red). (Box plot center-line: median; box limits: first and third quartile; whiskers: min/max values; Indicated *P*-values calculated from unpaired two-tailed Students t-test; Numbers of samples indicated on x-axis under respective class). Image analysis has been done together with Friedrich Preußer.



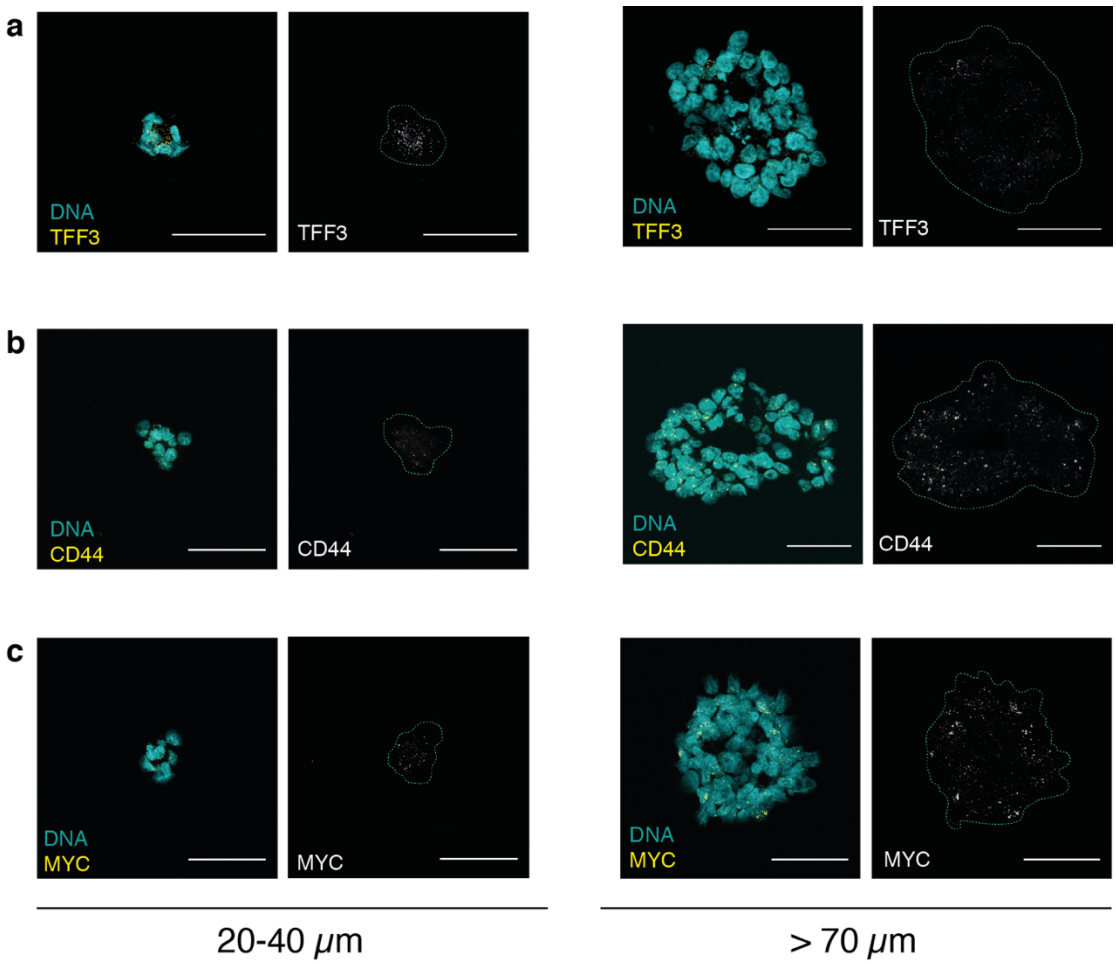
Supplementary Figure 3 I Detailed high-throughput pheno-seq workflow. (a) After staining and recovery (optional: fixation by DSP), spheroids are distributed into a nanowell chip by a microsolenoid-valve dispenser in 50 nl per well. For improvement of imaging quality, spheroids are centrifuged upside-down to the imaging foil and automatically imaged by inverted confocal laser scanning microscope. The chip is frozen at -80°C until further processing. **(b)** Images are processed by a custom-made image processing pipeline in KNIME and ImageJ. A Shiny-based web-app (PhenoSelect) enables interactive visualization, analysis and selection of spheroids based on quantified image features. **(c)** A filter-file generated in the PhenoSelect web-app is used to dispense RT/Amp reagents only in selected nanowells. cDNA generation and amplification are performed in the nanowell chip. After pooling of barcoded cDNA, 3´-library generation and next generation sequencing, resulting raw data can be de-multiplexed using internal barcodes listed in the welllist/feature-file generated by PhenoSelect. **(d)** Combined imaging and gene expression profiling enables combined analysis of gene expression and image features. Image processing pipeline and PhenoSelect has been developed together with Friedrich Preußer.



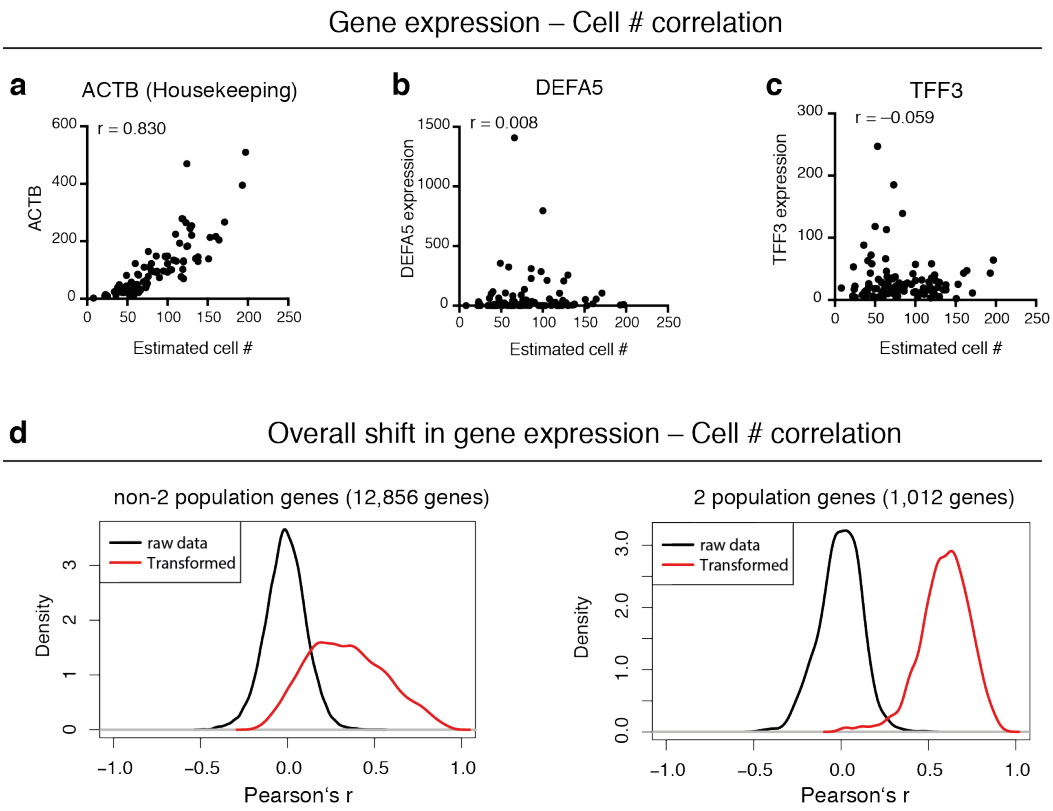
Supplementary Figure 4 I PhenoSelect software for interactive visualization, analysis and selection of spheroids from high-throughput pheno-seq. (a) Primary selection of single spheroids as well as analysis and curation of selected spheroids. Filter- and welllist/feature-files are generated at this point and can be reloaded or changed at any time. **(b)** PhenoSelect enables import of externally generated 2D tSNE maps (PAGODA) for combined analysis of image features and gene expression. PhenoSelect has been developed together with Friedrich Preußer.



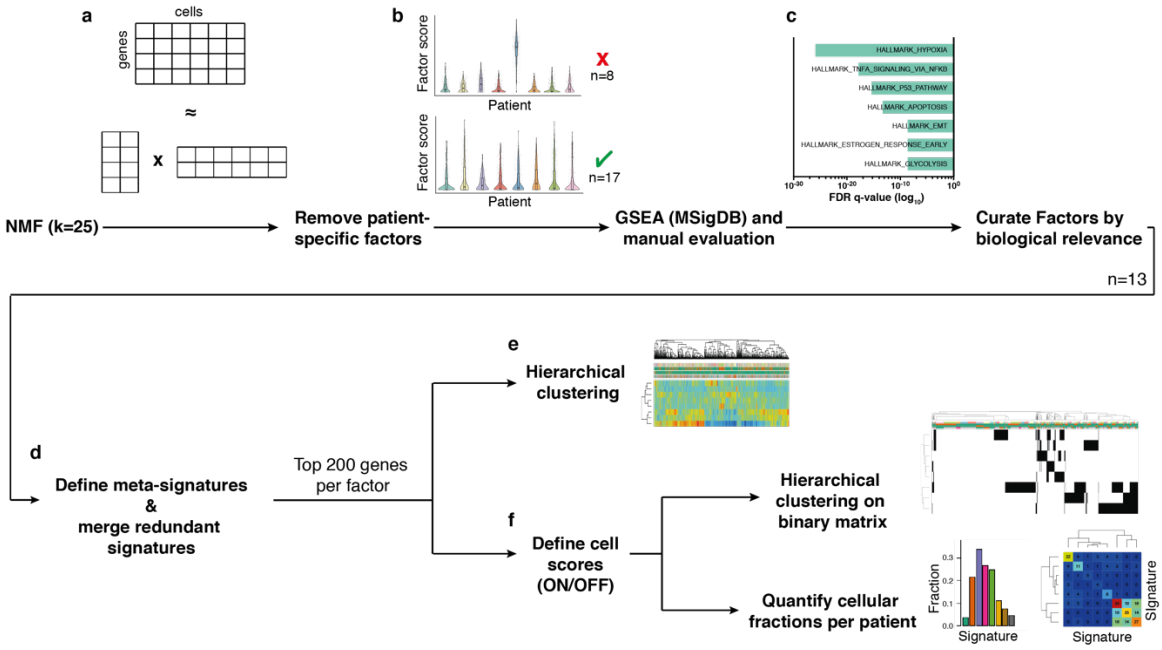
Supplementary Figure 5 I HT-pheno-seq cluster specific image features and comparison to manual pheno-seq. (a and b) Same t-SNE map for MCF10CA pheno-seq data as shown in Figure 2.5, but colored for image features 'size' (a) and 'skewness' (b). Right: Violin plots reflect image feature quantification per cluster (k-means clustering: k=2; violin center-line: median; box limits: first and third quartile; whiskers: ±1.5 IQR; Indicated *P*-values from unpaired two-tailed Students t-test). Image feature associations can be interpreted according to the biological background (e.g. proliferation and cell density). **(c)** Venn-diagrams showing overlaps of identified spheroid phenotype-specific genes between manual pheno-seq and HT-pheno-seq based on differential expression analysis (fold change > 1.3; adjusted p-value < 0.1). RNA-seq analysis has been performed together with Jeongbin Park.



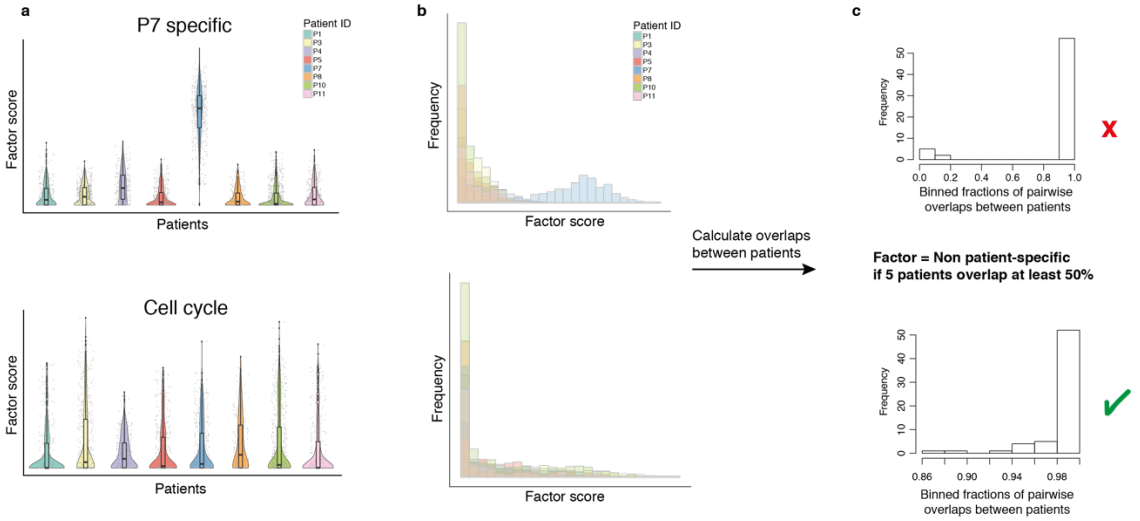
Supplementary Figure 6 I Validation of CRC pheno-seq data by RNA-FISH. (a, b, c) Representative RNA-FISH images (Z-projections) of different spheroid size classes for differentiation marker TFF3 (a) and cancer stem cell markers CD44 (b) and MYC (c). RNA-FISH staining of big (>70 μm) and small (20-40 μm) spheroids with (left) and without (right) Hoechst counterstain visualization (Hoechst: cyan; RNA: yellow). Dashed line in images without Hoechst visualization represents spheroid border (scale bar 50 μm).



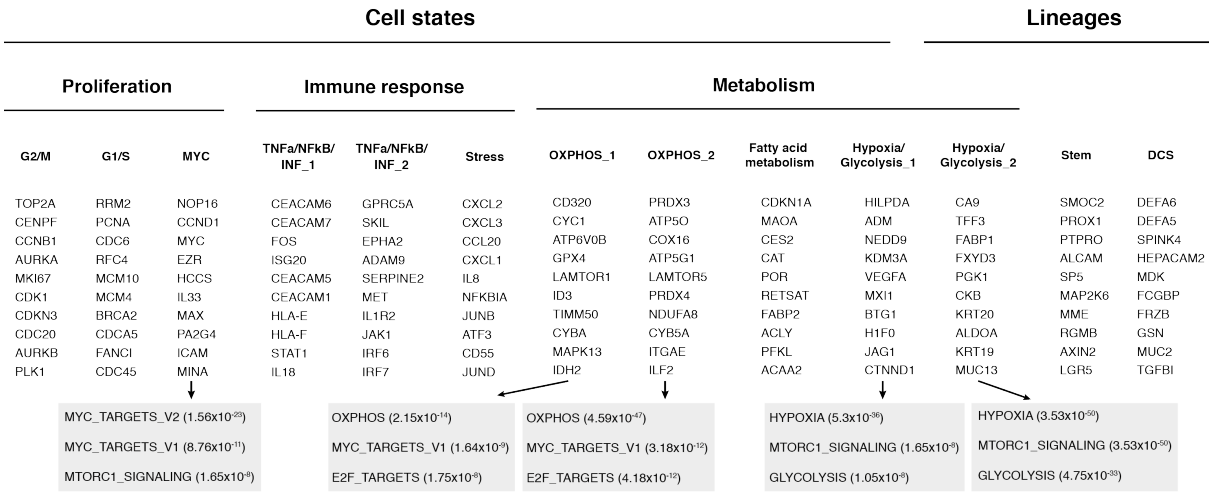
Supplementary Figure 7 | Gene-specific and global correlation analysis of expression with estimated cell numbers after CRC pheno-seq data transformation. **(a)** Relations of estimated cell numbers and downsampled mRNA counts visualized as scatter plots as well as associated Pearson's correlation coefficients (r) for housekeeping gene ACTB and differentiation markers TFF3 and DEFA5. **(b)** Pearson's correlation coefficient (r) distributions of gene expression and cell numbers for all genes before and after data transformation (cell number dependent downsampling) subdivided into non-2 population genes (left) and 2-population genes (right) as identified by maximum likelihood inference. RNA-seq analysis in has been performed together with Christiane Fuchs and Lisa Amrhein.



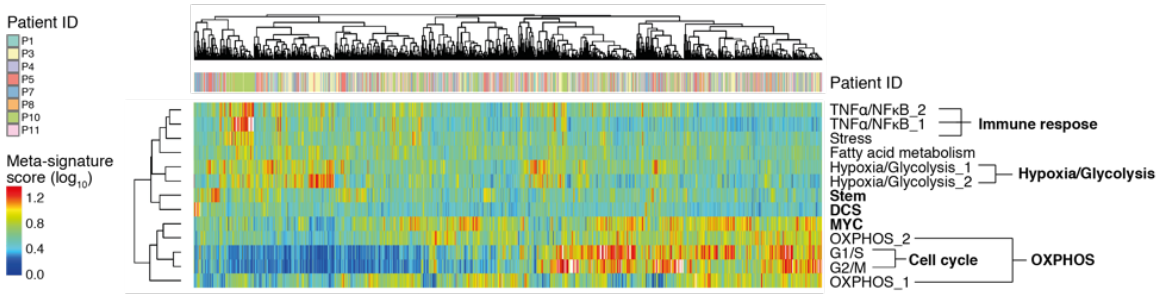
Supplementary Figure 8 | Workflow to identify shared expression programs across single cells of CRC patients by non-negative matrix factorization (NNMF). (a) NNMF is applied to the normalized and mean-centered single cell expression matrix of eight LGR5+ CRC patients with predicted factor numbers of k=25. **(b)** Patient-specific factors are removed by calculating overlaps in factor score distributions between patients. **(c)** Top 200 genes per factor are evaluated for biological relevance by GSEA and manual curation. **(d)** Meta-signature scores per cell are defined by averaged expression of top 200 genes per factor and redundant factors with similar enrichments and clustering are merged, resulting in eight core meta-signatures. **(e)** Core meta-signature scores can be clustered to evaluate relations between identified factors. **(f)** Binary ON/OFF states of core meta-signatures are inferred per cell that can be used for clustering or to quantify cellular fractions per patient and to identify cells with multiple active signatures per patient. The NNMF workflow was jointly developed with Teresa Krieger.



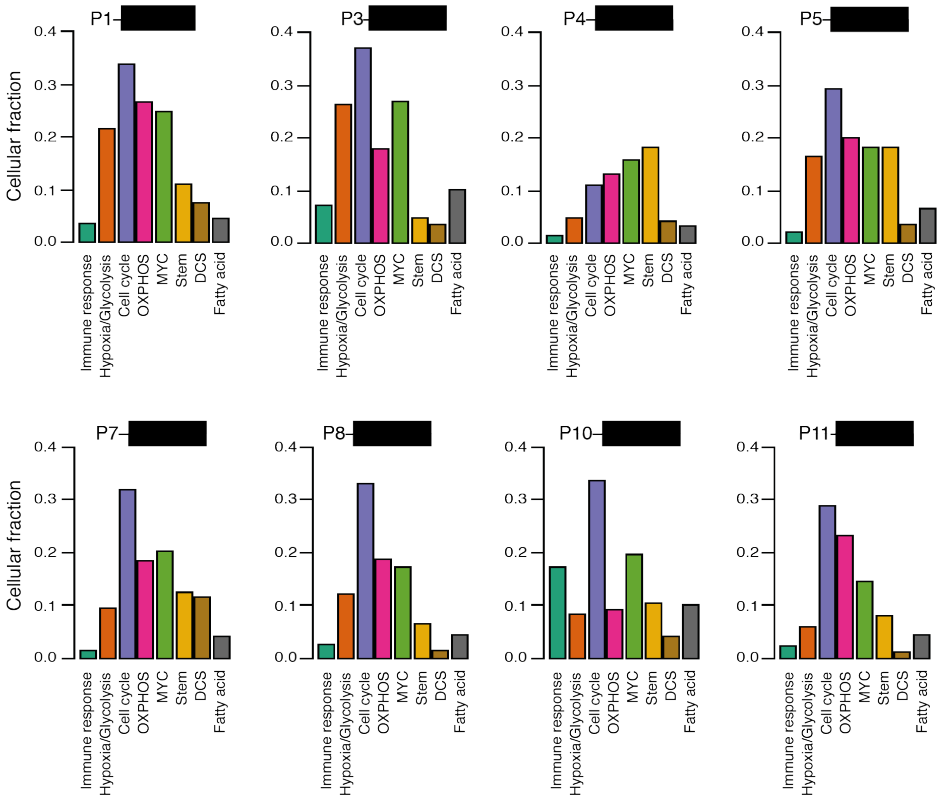
Supplementary Figure 9 | Identification of patient-specific factors identified by NNMF. (a) Violin plots of factor scores plotted per patient. Shown are patient specific (upper) and shared gene expression program (lower: cell cycle). **(b)** Frequency distributions of factor scores per patient for factors shown in (a). **(c)** Fractions of pairwise factor score overlaps between patients plotted as binned frequency distribution. Factors are removed from further analysis if less that 5 patients overlap at least 50%. The NNMF workflow was jointly developed with Teresa Krieger.



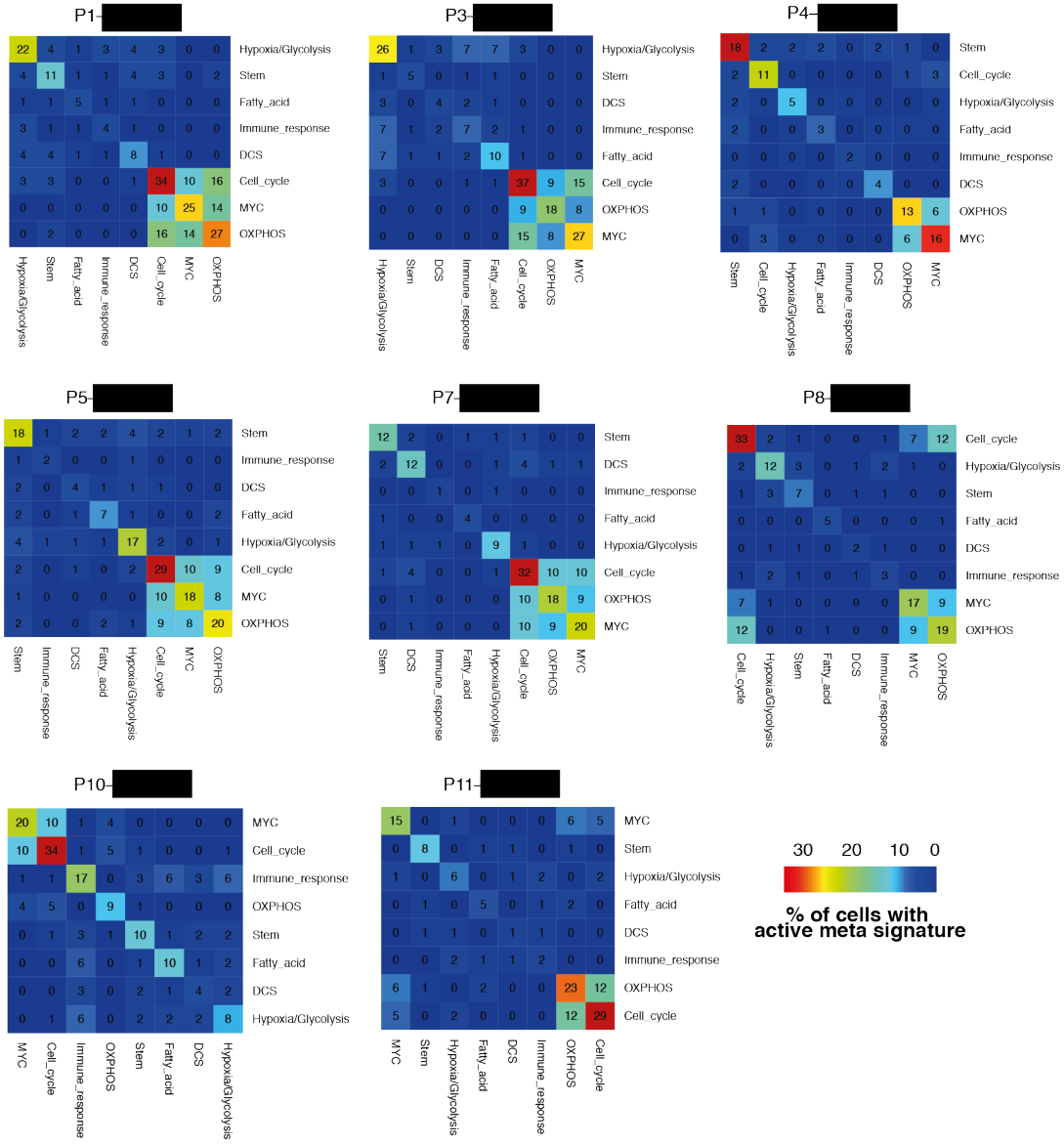
Supplementary Figure 10 | Biological classification of factors identified by NNMF derived from LGR5+ CRC patients. Identified factors (n=13) and associated top 200 genes can be categorized according to their biological enrichments (e.g. G2/M or OXPHOS). These can be classified in four main categories: cell cycle, metabolism, differentiation and immune response. Shown are 10 representative genes per factor ranked by NNMF factor score. Grey boxes below contain major gene set enrichments of HALLMARK gene sets[231] for representative factors.



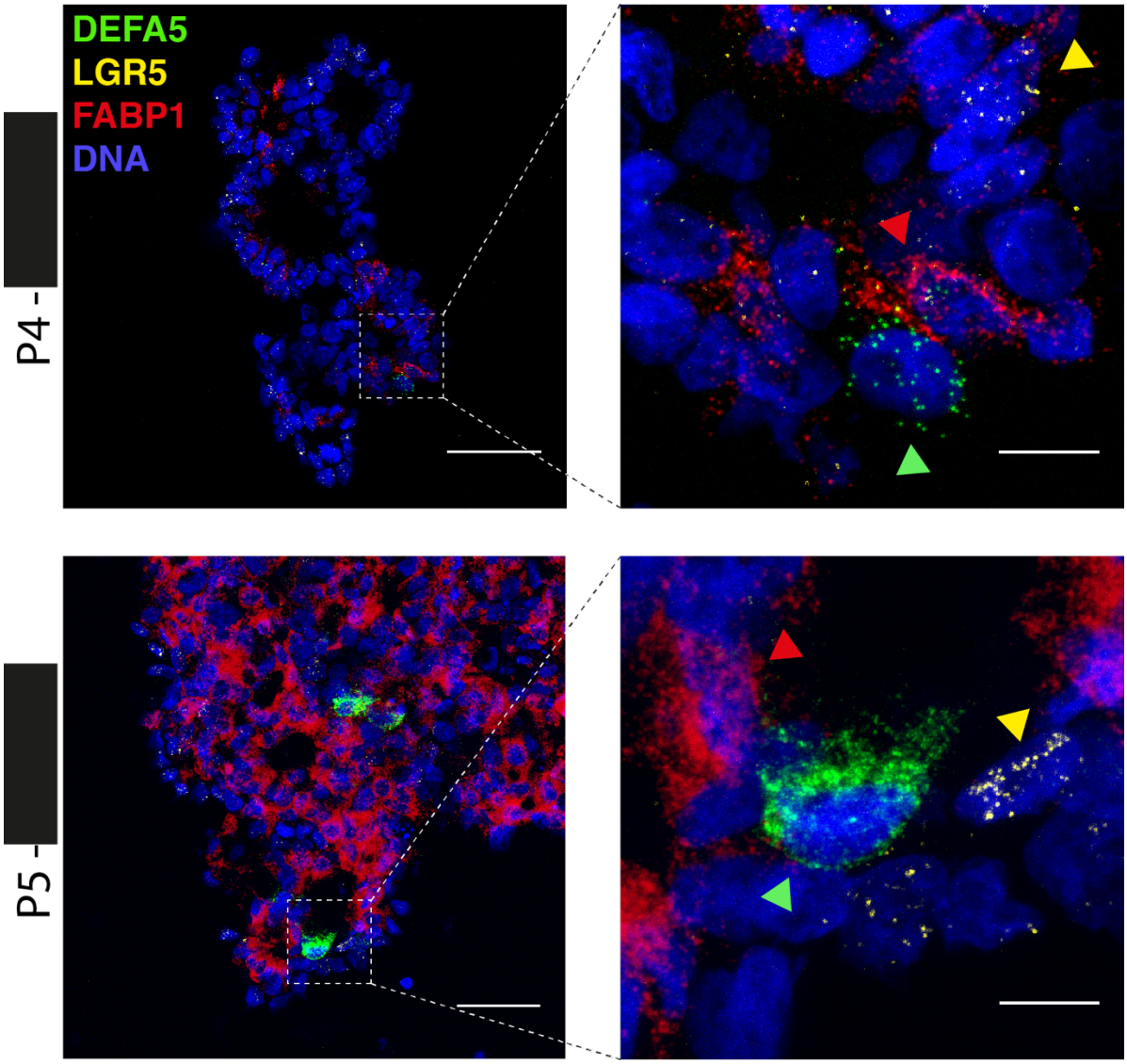
Supplementary Figure 11 | Clustering of meta-signature scores identified by NNMF. Hierarchical clustering of residual 13 meta-signature scores after removal of patient-specific and non-relevant factors. Dendrograms reflect overall clustering and row below shows patient ID information. Names of meta-signatures reflecting enriched genes are listed beside the heatmap. Core-signatures are written in bold letters and can originate from merged meta-signatures. Data analysis was jointly performed with Teresa Krieger.



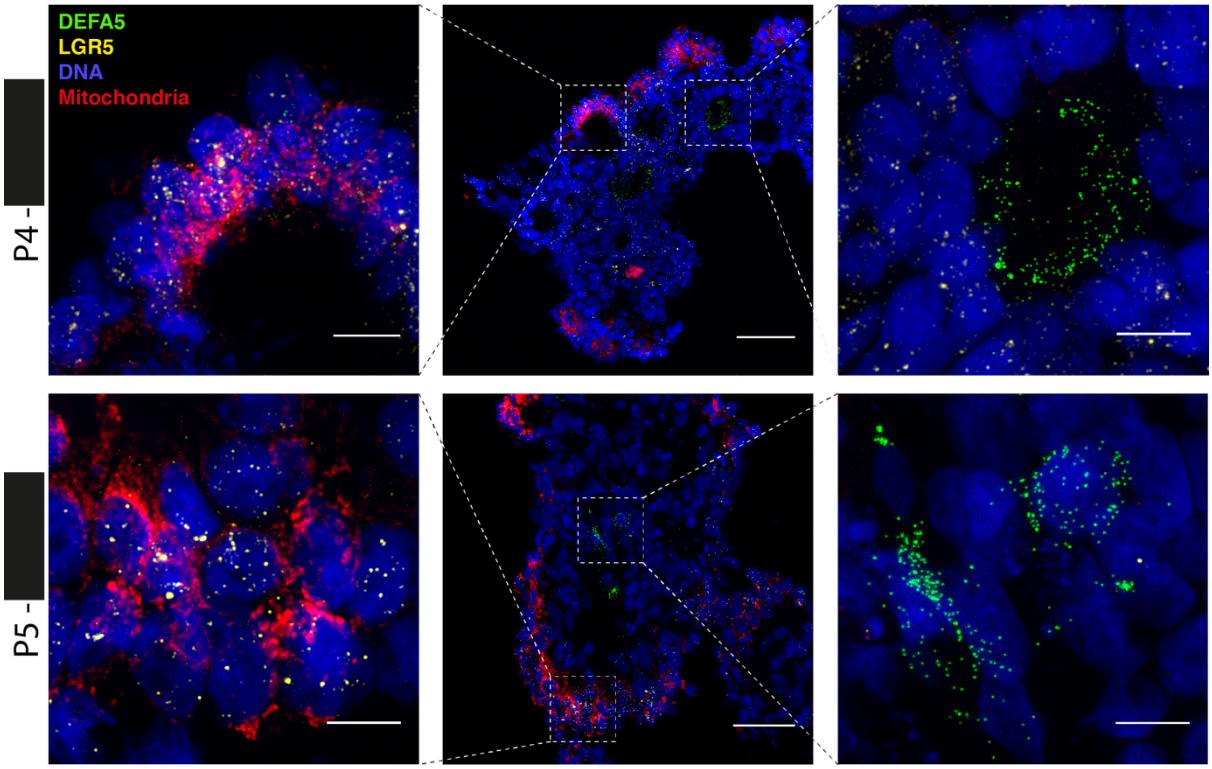
Supplementary Figure 12 | Cellular fractions of active core meta-signatures vary between patients. Barplots showing cellular fractions with active core meta-signature (reflecting ON/OFF states of respective signatures per cell) per patient identified by NNMF. Data analysis was jointly performed with Teresa Krieger.



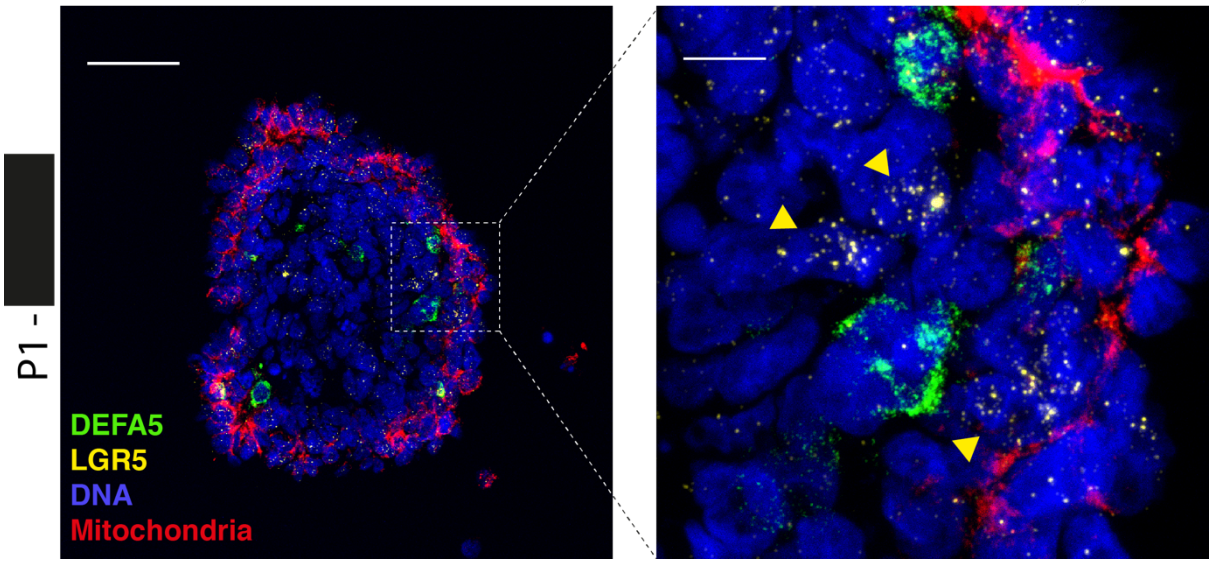
Supplementary Figure 13 | Identification of cells with multiple active signatures per patient. Heatmaps showing fractions of cells with active meta-signatures per patient (diagonal from upper left to lower right) and fractions of cells with pairwise overlaps of meta-signatures. Data analysis was jointly performed with Teresa Krieger.



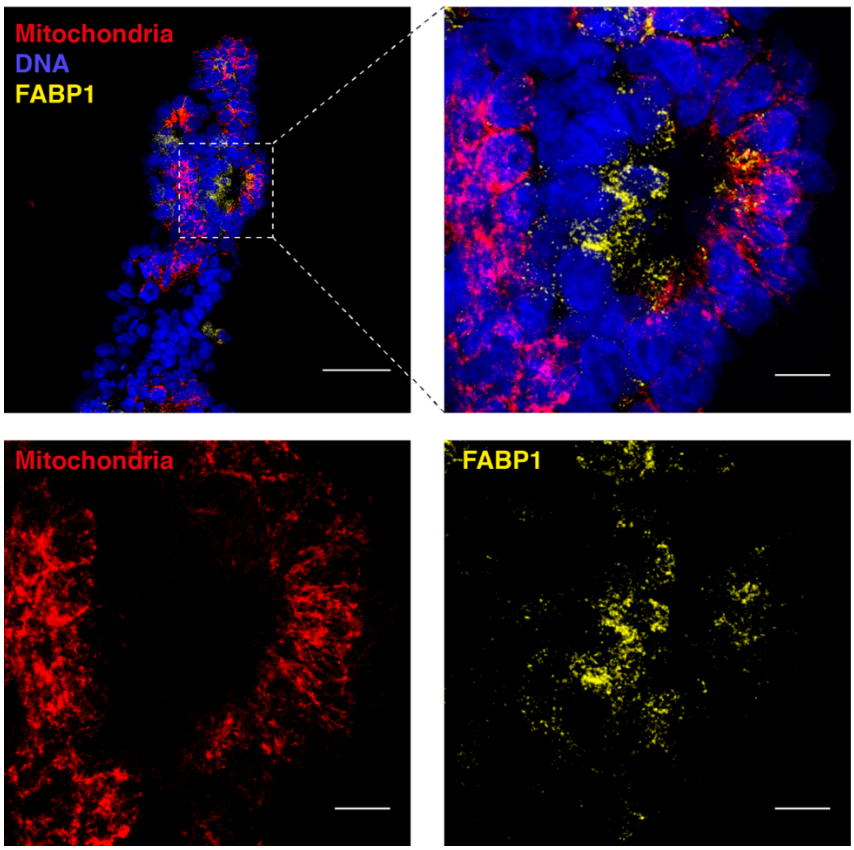
Supplementary Figure 14 | Validation of lineage-specific marker genes by RNA-FISH. Histological sections of CRC spheroids derived from P4 (upper) and P5 (lower) co-stained for representative lineage-specific marker genes by RNA-FISH. Left: overview images; scale bar 50 μm. Right: magnified images that represent dashed box regions in overview images (4x digital zoom); scale bar 10 μm. DEFA5: Paneth cells (green), LGR5: stem cells (yellow), FABP1: differentiated cells (red). Colored arrowheads mark associated subtypes in magnified images. Images represent Z-projections from 10 μm slices. DNA counterstain by DAPI (blue).



Supplementary Figure 15 | Spatial location of LGR5⁺ cells coincides with mitochondrial abundance in budding regions. Histological sections of CRC spheroids derived from P4 (upper) and P5 (lower) co-stained for representative lineage-specific marker genes by RNA-FISH and for mitochondria with Mitotracker Red CMXRos (100 nM). Middle: overview images; scale bars 50 μm. Left and right: magnified images that represent dashed box regions in overview image (4x digital zoom); scale bars 10 μm. DEFA5: Paneth cells (green), LGR5: stem cells (yellow), mitochondria (red). Images represent Z-projections from 10 μm slices. DNA counterstain by DAPI (blue).

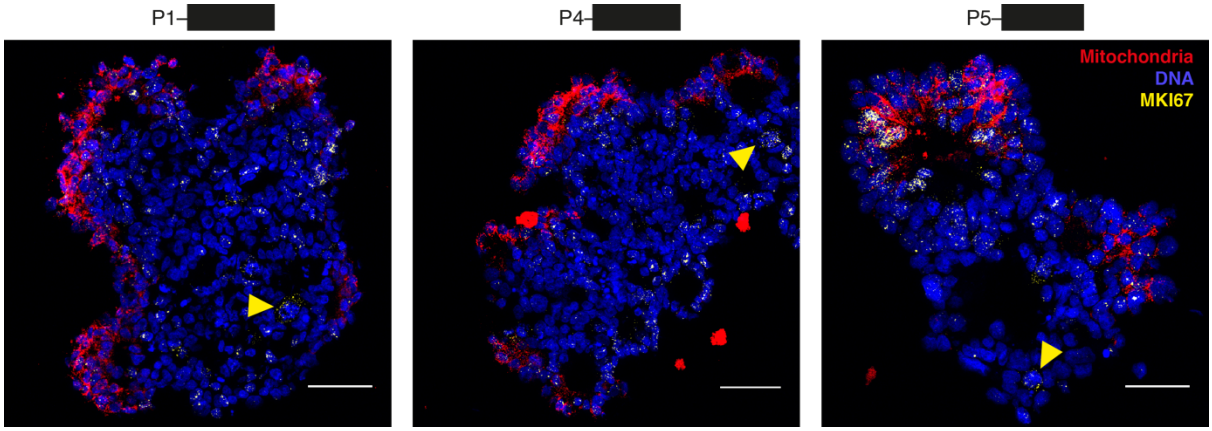


Supplementary Figure 16 | LGR5⁺ cells are not restricted to the outer layer of CRC spheroids. Histological section of CRC spheroid derived from P1 co-stained for representative lineage-specific marker genes by RNA-FISH and for mitochondria with Mitotracker Red CMXRos (100 nM). Left: overview image; scale bar 50 μm. Right: magnified image that represents dashed box region in overview image (4x digital zoom); scale bar 10 μm. DEFA5: Paneth cells (green), LGR5: stem cells (yellow), mitochondria (red). Yellow arrowheads mark LGR5⁺ cells. Images represent Z-projections from 10 μm slices. DNA counterstain by DAPI (blue).

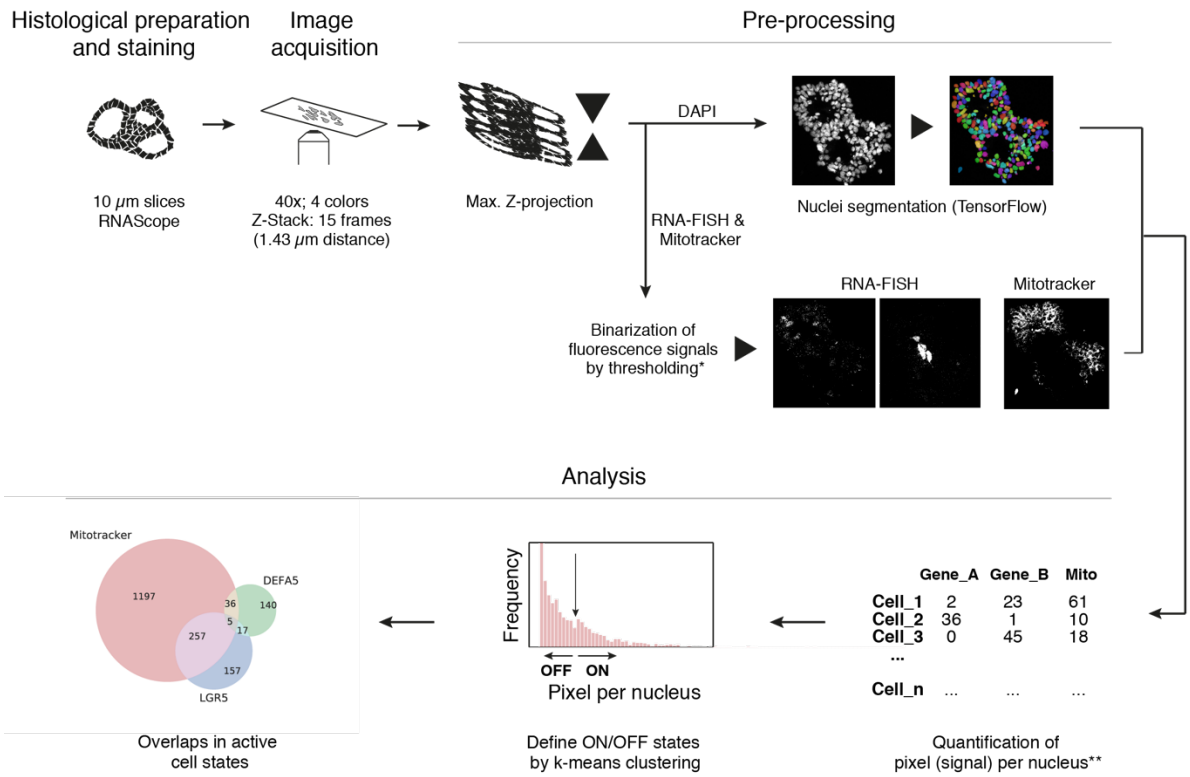


Supplementary Figure 17 | Differentiated FABP1^{+} cells are primarily located in OXPHOS^{Low} regions.

Histological section of CRC spheroid derived from P4 co-stained for lineage-specific marker gene FABP1 (yellow) by RNA-FISH and for mitochondria (red) with Mitotracker Red CMXRos (100 nM). Top left: overview image; scale bar 50 μm. Top right: magnified image that represents dashed box region in overview image (4x digital zoom); scale bar 10 μm. Lower left and right: Individual channels of magnified image. Images represent Z-projections from 10 μm slices. DNA counterstain by DAPI (blue).



Supplementary Figure 18 | Actively cycling cells frequently locate to OXPHOS^{high} regions. Histological sections of CRC spheroids derived from P1, P4 and P5 co-stained for cell cycle marker MKI67 (yellow) by RNA-FISH and for mitochondria (red) with Mitotracker Red CMXRos (100 nM). Scale bars 50 μm. Images represent Z-projections from 10 μm slices. DNA counterstain by DAPI (blue).



Supplementary Figure 19 | Methodological overview for quantitative *in situ* analysis of CRC spheroids by RNA-FISH and Mitotracker. After Mitotracker staining, fixation and embedding, CRC spheroids underwent histological preparation and RNA-FISH staining (RNAScope). Images were acquired semi-automatically on a confocal laser scanning microscope (Leica SP8). Image pre-processing involved nuclei segmentation by deep-learning (TensorFlow) as well as thresholding for binarization of fluorescence signals (*RNA-FISH: maximum entropy, Mitotracker: moments-preservation). Per channel, pixels were then counted per nucleus (**for Mitotracker quantification, nuclei were expanded by morphological dilation to acquire cytoplasmic signals from mitochondria) and ON/OFF states per cell for each channel were defined by k-means clustering on frequency distributions of pixel counts per nucleus. Finally, overlaps of active cell states can be computed. The image analysis pipeline was developed together with Foo Wei Ten.

6.2 Supplementary Tables

Supplementary Table 1 | Dataset overview and QC metrics for pheno-seq project

3D-culture model	MCF10CA	MCF10CA	MCF10CA	MCF10CA	MCF10CA	CRC spheroid
Method	scRNA-seq	Pseudo pheno-seq	M-pheno-seq	HT-pheno-seq (bottom control)	HT-pheno-seq (DSP)	HT-pheno-seq
Library structure	Full-length C1	Full-length C1	Full-length Tube-based	3`-end iCELL8	3`-end iCELL8	3`-end iCELL8
Sequencer	HiSeq 2000	HiSeq 2000	HiSeq 2000	HiSeq 2000	NextSeq 500	NextSeq 500
Sample # (after QC)	166	8	8	64	210	95
Origin	Cell line	Cell line	Cell line	Cell line	Cell line	Patient-derived
Mean total read count per sample	3,820,057	3,685,536	9,965,986	485,975	803,669	1,304,480
Mean detected genes (> 0) per sample (all reads)	8,844	15,783	12,360	8,458	8,221	9,891
Mean detected genes (> 0) per sample (down-sampled to 100k reads)	5,554	13,374	8,411	7,051	6,377	7,412

Supplementary Table 2 | Diagnostic background of CRC spheroid cultures*

Patient #	ID	Sex	Year of birth	Diagnosis	Origin
P1	[redacted]	male	[redacted]	[redacted]	Liver metastasis
P2		male			Lung metastasis
P3		female			Liver metastasis
P4		female			Liver metastasis
P5		female			Primary tumor
P6		female			Primary tumor
P7		male			Liver metastasis
P8		male			Liver metastasis
P9		male			Primary tumor
P10		n.a.			Primary tumor
P11		n.a.			Primary tumor
P12		male			Primary tumor

(*Information provided by Dr. Claudia Ball/Prof. Dr. Hanno Glimm)

Supplementary Table 3 | Phenotypic overview of CRC spheroid cultures

Patient #	ID	Spheroid phenotype	Growth rate	Culture time	Dissociation time	Validation assays
P1		compact	++	9-10 days	30 min	✓
P2		loose connection	+++	6-7 days	20 min	✗
P3		compact	+	14 days	30 min	✗
P4		compact	+++	6-7 days	20 min	✓
P5		compact	+++	6-7 days	30 min	✓
P6		loose connection, Mucus	+++	6-7 days	15 min	✗
P7		compact	++	9-10 days	30 min	✗
P8		compact	++	9-10 days	30 min	✗
P9		loose connection	+++	6-7 days	10 min	✗
P10		loose connection	+++	6-7 days	25 min	✗
P11		compact/loose	+++	6-7 days	35 min	✗
P12		compact/loose	+++	6-7 days	25 min	✗

Supplementary Table 4 | Media, supplements, buffers and chemicals

Product	Application	Company
Phosphate buffered saline (PBS)	Buffer	Sigma
DMEM/F12	Cell culture medium	Life Technologies
Advanced DMEM/F12	Cell culture medium	Life Technologies
Fluorobrite DMEM	Cell culture medium	Life Technologies
DMEM	Cell culture medium	Life Technologies
Fetal bovine serum	Culture medium supplement	Life Technologies
Bovine serum albumin	Culture medium supplement	Life Technologies
Horse Serum	Culture medium supplement	Life Technologies
HEPES	Culture medium supplement	Sigma
Epidermal growth factor (EGF)	Culture medium supplement	R&D Systems
Fibroblast growth factor (FGF)	Culture medium supplement	R&D Systems
D-Glucose	Culture medium supplement	Sigma
L-Glutamine	Culture medium supplement	Sigma
Heparin	Culture medium supplement	Sigma
Cholera Toxin	Culture medium supplement	Sigma
Insulin	Culture medium supplement	Sigma
Hydrocortisone	Culture medium supplement	Sigma
Matrigel	ECM surrogate	Corning
Glycine	Immunofluorescence	Sigma
Tween20	Immunofluorescence	Sigma
NaN_3	Immunofluorescence	Sigma
Goat serum	Immunofluorescence	Sigma
Dispase	ECM dissociation	Sigma
Accumax	Cellular dissociation	StemCell Technologies
Trypsin 0.05% & 0.25%	Cellular dissociation	Life Technologies
Formaldehyde solution (Methanol-free)	Cellular fixation	Thermo Fisher
dithio- bis(succinimidyl propionate) (DSP)	Cellular fixation	Sigma
Richard-Allan Scientific Neg-50 Frozen Section Medium	Histology	Thermo Fisher
Sucrose	Histology	Sigma

SlowFade Gold Antifade solution	Histology	Thermo Fisher
fluorescein sodium salt	Fluorescence microscopy	Sigma
PF-03084014	γ-secretase inhibitor	Sigma

Supplementary Table 5 | Cell culture plates and flasks

Plates/Flask type	Company
25 cm^2 culture flasks	greiner
24-well cell culture plates	greiner
15μ angiogenesis slides	ibidi
8-well Nunc Lab-Tek Chamber Slides	Thermo Fisher
GravityTRAP ultra-low attachment 96-well plates	InSphero
75 cm^2 ultra-low attachment flasks	Corning
60 mm Ultra Low Attachment Culture Dishes	Corning
Aggrewell400 6-Well plates	StemCell Technologies

Supplementary Table 6 | PCR cycler programs for RT and cDNA amplification

SMARTer C1 (scRNA-seq)		Smart-Seq v4 (M-pheno-seq)		HT-pheno-seq & CRC scRNA-seq	
72°C – 3 min		72°C – 3 min		50°C – 3 min	
4°C – 10 min		4°C – ∞		4°C – 5 min	
25°C – 1 min				42°C – 90 min	
4°C – ∞		42°C – 90 min		50°C – 2 min	2 cycles
		70°C – 10 min		42°C – 2 min	
42°C – 90 min		4°C – ∞		70°C – 15 min	
70°C – 10 min				95°C – 1 min	
4°C – ∞		95°C – 1 min		98°C – 10 sec	
		98°C – 10 sec		65°C – 30 sec	18 cycles
95°C – 1 min		65°C – 30 sec	16 cycles	68°C – 3 min	
95°C – 20 sec		68°C – 3 min		72°C – 10 min	
58°C – 4 min	5 cycles	72°C – 10 min		4°C – ∞	
68°C – 6 min		4°C – ∞			
95°C – 20 sec					
64°C – 30 sec	9 cycles				
68°C – 6 min					
95°C – 30 sec					
64°C – 30 sec	7 cycles				
68°C – 7 min					
72°C – 10 min					
4°C – ∞					

Supplementary Table 7 | Kits and reagents for reverse transcription, cDNA amplification and sequencing library preparation

Application	Protocol	Library structure	Component	Company	Pheno-seq	CRC scRNA-seq
scRNA-seq	SMARTer C1	Full-length	SMARTer kit for C1	TakaraBio	✓	✗
			C1 IFC for mRNA-seq 10-17 μm	Fluidigm	✓	✗
			Nextera XT kit	Illumina	✓	✗
			Ampure XP beads	Beckman Coulter	✓	✗
Manual pheno-seq	Smart-Seq v4	Full-length	Smart-Seq v4 kit	TakaraBio	✓	✗
			PicoPure RNA-isolation kit	Life-technologies	✓	✗
			Nextera XT kit	Illumina	✓	✗
			Ampure XP beads	Beckman Coulter	✓	✗
HT-pheno-seq & scRNA-seq	Rapid development	3´-end	SmartChip v1/v2 kit	TakaraBio	✓	✓
			Recombinant RNase Inhibitor	TakaraBio	✓	✓
			DNA Clean and Concentrator–5 kit	Zymo Research	✓	✓
			Nextera XT kit	Illumina	✓	✓
			Betaine	Sigma	✓	✓
			dTNPs	TakaraBio	✓	✓
			$MgCl_2$	Invitrogen	✓	✓
			Dithiothreitol	TakaraBio	✓	✓
			5x SMARTScribe™ first-strand buffer	TakaraBio	✓	✓
			2x SeqAmp™ PCR buffer	TakaraBio	✓	✓
			Triton X-100	Acros	✓	✓
			SMARTScribe Reverse Transcriptase	TakaraBio	✓	✓
			SeqAmp DNA Polymerase	TakaraBio	✓	✓
			Ampure XP beads	Beckman Coulter	✓	✓

Supplementary Table 8 I Microscopes and objectives

Microscope	Objective	Application	Pheno-seq	CRC scRNA-seq
Olympus BX43 (iCELL8)	4x Air (iCELL8 in-built)	iCELL8	✓	✓
Leica SP8	40x/1.30 oil (Leica HC APO CS2)	RNA-FISH	✓	✓
	10x/0.30 air objective (Leica HC PL FLUOTAR)	Single cell seeding	✓	✗
		γ-secretase inhibitor assay	✓	✗
		HT-pheno-seq	✓	✗
		Leakage test	✓	✗
Zeiss Axio observer	10x/0.3 air (Zeiss EC PLAN-NEOFLUAR)	Reseeding assay	✓	✗
		Immunofluorescence	✓	✗
ASI diSPIM	40x/0.80 water (Nikon NIR-Apo)	Cell count reference	✓	✗

Supplementary Table 9 I Antibodies and live dyes

Target/Antigen	Antibody/dye #	Host	Concentration/ Dilution	Company
anti-Vimentin (Alexa Fluor 594)	EPR3776	rabbit	1:100	abcam
anti-β-Actin	8H10D10	mouse	1:200	Cell Signaling
anti-Cytokeratin 15	LHK15	mouse	1:50	ThermoFisher
anti-mouse (Alexa Fluor 594)	8890	goat	1:200	Cell Signaling
Mitochondria	Mitotracker Red CMXRos	-	100 nM	Thermo Fisher
Cytoplasm	CellTracker Red CMTPX	-	10 μM	Thermo Fisher
DNA	Hoechst 33258	-	1 μg/ml	Thermo Fisher
DNA and dead cells	Readyprobes Cell Viability Imaging Kit	-	n.a.	Invitrogen

Supplementary Table 10 | ACDbio RNAScope probes for RNA-FISH

Gene	Channel	Catalog #	Pheno-seq	Pheno-seq subtype	CRC scRNA-seq	CRC scRNA-seq Subtype
ATF3	C1	470861	✗		✓	Immune response
CCL20	C3	409611-C3	✗		✓	Immune response
CD44	C3	311271-C3	✓	big	✗	
CD9	C3	430671-C3	✓	big	✗	
CFB	C1	402101	✓	round	✗	
CKB	C1	480671	✗		✓	Glycolysis/Hypoxia
CLDN2	C1	492051	✗		✓	Stem
DEFA5	C1	423981	✓	Paneth	✓	Paneth
FABP1	C3	534801-C3	✗		✓	Glycolysis/Hypoxia
FTL	C1	315051	✓	small	✗	
HES1	C2	311191-C2	✗		✓	Stem
HES1	C1	311191	✗		✓	Stem
HES6	C2	521301-C2	✗		✓	Paneth
HILPDA	C1	320269/300031	✗		✓	Glycolysis/Hypoxia
JUN	C1	470541	✓	big	✓	Immune response
KDM3A	C1	454631	✗		✓	Glycolysis/Hypoxia
KRT18	C3	310211-C3	✓	small	✓	OXPHOS
KRT19	C3	426221-C3	✗		✓	Glycolysis/Hypoxia
LDHA	C1	487811	✗		✓	Glycolysis/Hypoxia
LDHB	C1	NM_001174097.2	✗		✓	OXPHOS
LGR5	C2	311021-C2	✗		✓	Stem
LGR5	C3	311021-C3	✗		✓	Stem
MKI67	C3	591771-C3	✗		✓	Cell Cycle
MYC	C2	311761-C2	✓	big	✓	MYC
PGK1	C1	310401	✗		✓	Glycolysis/Hypoxia
PROX1	C2	530241-C2	✓		✓	Stem
PROX1	C1	530241	✗	Stem	✓	Stem
RHOA	C1	416291	✓	aberrant	✗	
SNAI2	C1	554581	✓	aberrant	✗	
SOX4	C1	469911	✗		✓	Stem
SOX9	C1	404221	✗		✓	Stem
TFF3	C1	403101	✓	small	✓	Glycolysis/Hypoxia
VEGFA	C2	423161-C2	✗		✓	Glycolysis/Hypoxia

Supplementary Table 11 | Software and core extensions

Software	Components/packages	Reference
R / RStudio	Stats	https://www.r-project.org https://www.rstudio.com
	ggplot2	
	shiny	
	PAGODA/SCDE	
	ComplexHeatmaps	
	stochprofML2	
	Seurat	
In-house RNA-seq pipeline (Roddy based)	FastQC	https://github.com/TheRoddyWMS/Roddy
	Cutadapt	
	Star	
	featureCounts	
KNIME	KNIME Analytics Platform	https://www.knime.com
	KNIME File Handling Nodes	
	KNIME Image Processing	
	KNIME Image Processing - ImageJ Integration (Beta)	
	KNIME Interactive R Statistics Integration	
Microsoft Office 2016	Word	https://www.office.com
	Excel	
	Powerpoint	
Python	tensorflow	https://www.python.org https://www.tensorflow.org
	numpy	
	pandas	
	matplotlib	
	scikit-learn	
	scikit-image	
	OpenCV	
MATLAB	Non-negative matrix factorization (nnmf)	https://de.mathworks.com
GraphPad Prism 7		https://www.graphpad.com
Adobe Illustrator		https://www.adobe.com/de
Mendeley Desktop		https://www.mendeley.com

6.3 List of Figures

6.4 List of Tables

6.5 Abbreviations

Units

pg	picogram
ng	nanogram
μg	microgram
mg	milligram
nl	nanoliter
μl	microliter
ml	milliliter
nm	nanometer
μm	micrometer
mm	millimeter
cm	centimeter
M	molar
mol	mole
^{o}C	degree Celsius
sec	seconds
min	minutes
h	hours
rpm	revolutions per minute
g	relative centrifugal force
λ	wavelength

Further abbreviations

2D	two-dimensional
3D	three-dimensional
ACTB	Beta-Actin
Amp	amplification
APC	Adenomatosis Polyposis Coli Tumor Suppressor
ATAC-seq	Assay for Transposase-Accessible Chromatin using sequencing
ATP	adenosine triphosphate
BIC	bayesian information criterion
bp	base pair
BS-seq	bisulfite sequencing
BSA	bovine serum albumin
cDNA	complementary DNA
CLAHE	Contrast Limited Adaptive Histogram Equalization

CMS	consensus molecular subtypes
CNAs	copy number alterations
CNNs	convolutional neural networks
CNVs	copy number variations
CO_2	carbon dioxide
CPM	counts per million
CRC	colorectal cancer
CSC	cancer stem cell
CTCs	circulating tumor cells
cZ	corrected Z-score
Paneth	deep crypt secretory
DEFA5	Defensin A5
DNA	deoxyribonucleic acid
dNTP	deoxynucleotide triphosphate
DSG3	desmoglein 3
DSP	dithio-bis(succinimidyl) propionate
DTT	dithiothreitol
ECM	extracellular matrix
EGF	epidermal growth factor
EGFR	epidermal growth factor receptor
EMT	epithelial-mesenchymal transition’
F	frequency
FACS	fluorescence activated cell sorting
FAP	fibroblast activating protein
FEP	fluorinated ethylene propylene
FGF	fibroblast growth factor
FISH	fluorescence in situ hybridization
FPM	fragments per million
GLUL	Glutamate Ammonia Ligase
H1F0	histone Family Member 0
HC	high-content
HSCs	hematopoietic stem cells
HT	high-throughput
IF	immunofluorescence
IHC	immunohistochemistry
iPS	induced pluripotent stem cells
IQR	interquartile range
KRT	keratin
LCM	laser capture microdissection
LGR5	Leucin-rich repeat–containing G protein–coupled receptor 5
LN	log-normal
LSFM	light sheet fluorescence microscopy
MET	mesenchymal-epithelial transition
$MgCl_2$	magnesium chloride
mRNA	messenger RNA
MSI	microsatellite instable

MSigDB	Molecular Signature Database
NA	numerical aperture
NF-κB	nuclear factor kappa-light-chain-enhancer of activated B cells
NGS	next generation sequencing
NNMF	non-negative matrix factorization
OXPHOS	oxidative phosphorylation
P	patient
PAGODA	pathway and geneset overdispersion
PBS	phosphate buffered saline
PC	principal component
PCA	principal component analysis
PCR	polymerase chain reaction
poly(T)	poly thymine
Pop	population
qPCR	quantitative PCR
RNA	ribonucleic acid
ROS	reactive oxygen species
RT	reverse transcription
sc	single cell
seq	sequencing
sm	single molecule
TA	transit-amplifying
TCA	tricarboxylic
Tdiff	Terminally differentiated CRC subtype
TFF3	Trefoil Factor 3
TGF-b	Transforming Growth Factor Beta
TNF-a	Tumor Necrosis Factor alpha
TSCS	topographic single cell sequencing
tSNE	t-distributed stochastic neighbor embedding
UMIs	unique molecular identifiers
VIM	vimentin
WNT	Wingless int1